Data journalism as a liberal art

Amelia McNamara @AmeliaMN
University of St Thomas
www.amelia.mn

flickr: MattHagedorn

What are the liberal arts?

flickr: Allison

Liberal arts today can refer to academic subjects such as literature, philosophy, mathematics, and social and physical sciences; and liberal arts education can refer to overall studies in a liberal arts degree program. For both interpretations, the term generally refers to **matters not relating to the professional, vocational, or technical curriculum**.

Wikipedia, Liberal arts education

vo·ca·tion·al

/vōˈkāSH(ə)n(ə)l/

adjective: vocational

relating to an occupation or employment.

"they supervised prisoners in vocational activities"

- (of education or training) directed at a particular occupation and its skills.

"vocational school"

vo·ca·tion
/vōˈkāSH(ə)n/

noun: vocation; plural noun: vocations
a strong feeling of suitability for a particular career or occupation.
"not all of us have a vocation to be nurses or doctors"
synonyms: calling, life's work, mission, purpose, function, position, niche;

- a person's employment or main occupation, especially regarded as particularly worthy and requiring great dedication.
  "her vocation as a poet"
- a trade or profession.

Liberal arts schools — "I know them when I see them"

- ☑ Focused on undergraduate education
- ☑ Small (student body and class sizes)
- ☑ Desire for students to learn about a variety of fields
- ☑ Primarily granting BA degrees

flickr: Allison

My experience with the liberal arts

MACALESTER

B.A., English and mathematics

SMITH COLLEGE

3 years MassMutual Faculty Fellow and visiting assistant professor in Statistical and Data Sciences

UNIVERSITY of ST. THOMAS
MINNESOTA

Assistant professor of
Computer & Information Sciences

What is

science?

flickr: jannekestaaks

“Key concepts required to develop data acumen include mathematical foundations, computational foundations, statistical foundations, data management and curation, data description and visualization, data modeling and assessment, workflow and reproducibility, communication, domain-specific considerations, and ethical problem solving.”

–Data Science for Undergraduates: Opportunities and Options
National Academies, 2018

“Key Competencies for an undergraduate Data Science Major

Computational and Statistical Thinking
Mathematical Foundations
Model Building and Assessment
Algorithms and Software Foundation
Data Curation
Knowledge Transference – **Communication** and Responsibility”

–Curriculum Guidelines for Undergraduate Programs in Data Science
Park City Math Institute (PCMI)

“Recommendation 2.1: Academic institutions should embrace **data science** as a vital new field that **requires specifically tailored instruction** delivered through majors and minors in data science as well as the development of a cadre of faculty equipped to teach in this new field.”

"As instructors rework individual classes based on outcomes and evaluation, it is likely that they will replace borrowed content from existing courses with **original materials** that fit together more naturally and better match personal educational styles or the culture of that institution or department."

–Data Science for Undergraduates: Opportunities and Options
National Academies, 2018

“Most institutions will implement a Data Science major from current courses in existing disciplines, **perhaps transitioning to more fully integrated courses** as outlined in the Appendix at a future date.”

"6.4. Related Courses

- Introduction to [Partner Discipline]
- Intermediate course in Discipline
- **Capstone Course with Data Experience and Projects**
- Two courses in writing, preferably one in technical writing.
- Public Speaking
- Ethics

[…] highlighed courses cover the bare necessities of the material required for a Data Science major"

–Curriculum Guidelines for Undergraduate Programs in Data Science
Park City Math Institute (PCMI)

Communicating data

Visualizing data— often exists in standalone courses

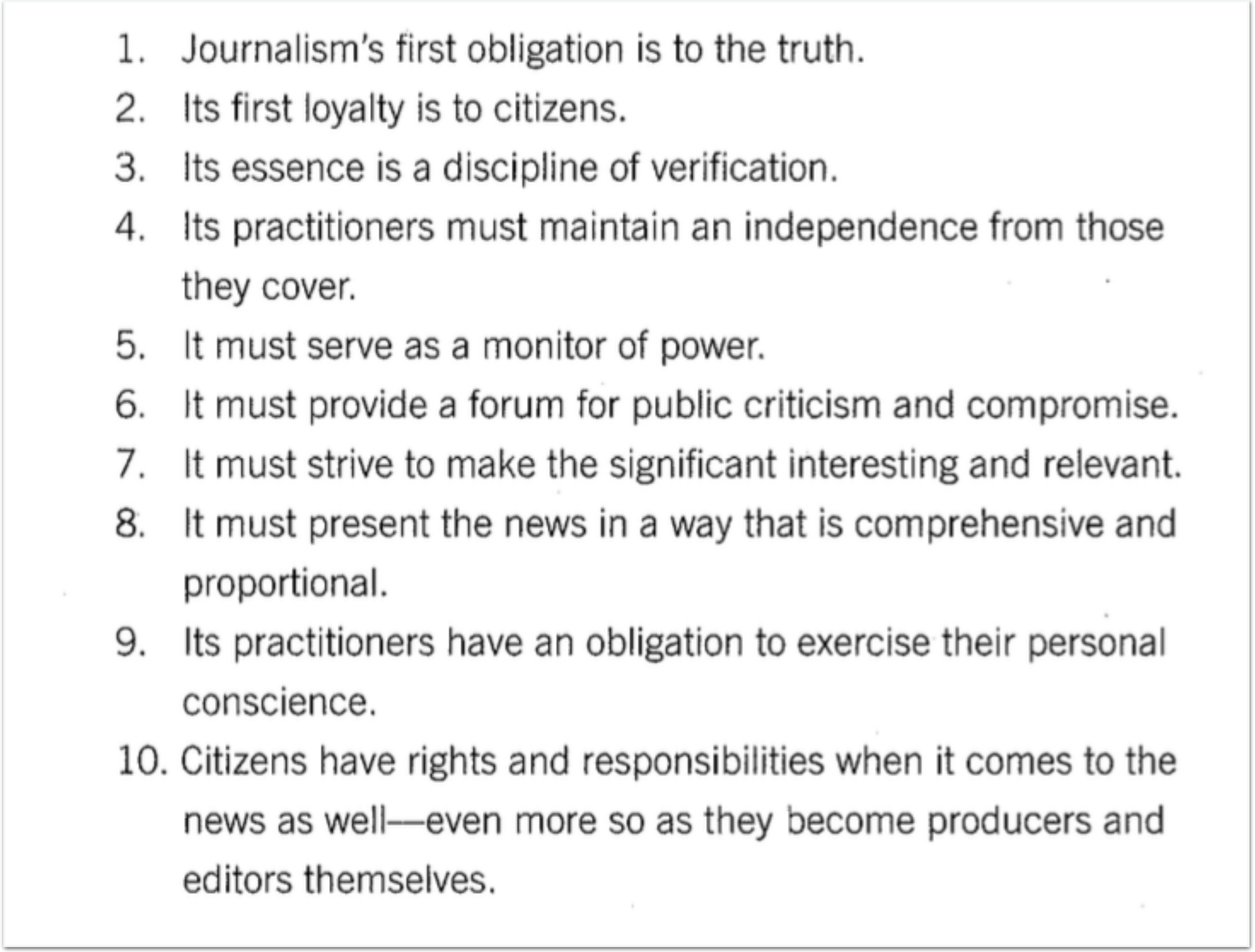
Writing data— typically integrated into other courses (e.g., regression modeling) or outsourced to other departments

Speaking data— rarely taught

flickr: lidor

What is journalism?

flickr: jdhancock



1. Journalism's first obligation is to the truth.
2. Its first loyalty is to citizens.
3. Its essence is a discipline of verification.
4. Its practitioners must maintain an independence from those they cover.
5. It must serve as a monitor of power.
6. It must provide a forum for public criticism and compromise.
7. It must strive to make the significant interesting and relevant.
8. It must present the news in a way that is comprehensive and proportional.
9. Its practitioners have an obligation to exercise their personal conscience.
10. Citizens have rights and responsibilities when it comes to the news as well—even more so as they become producers and editors themselves.

From The Elements of Journalism,
Bill Kovach and Tom Rosenstiel

Newsworthiness

- Timing
- Significance
- Proximity
- Prominence
- Human interest

https://www.mediacollege.com/journalism/news/newsworthy.html

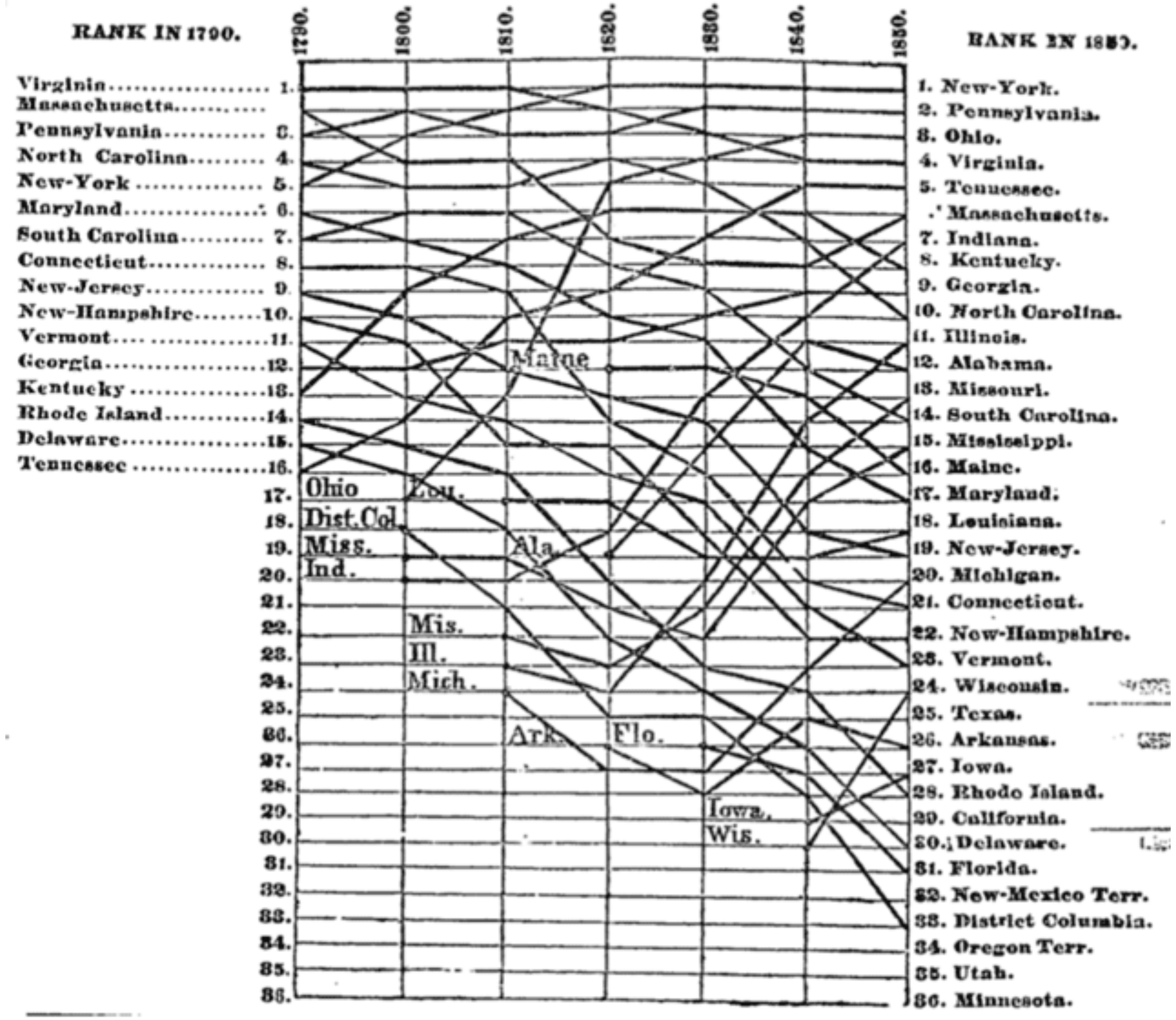
What is data journalism?

flickr: jdhancock

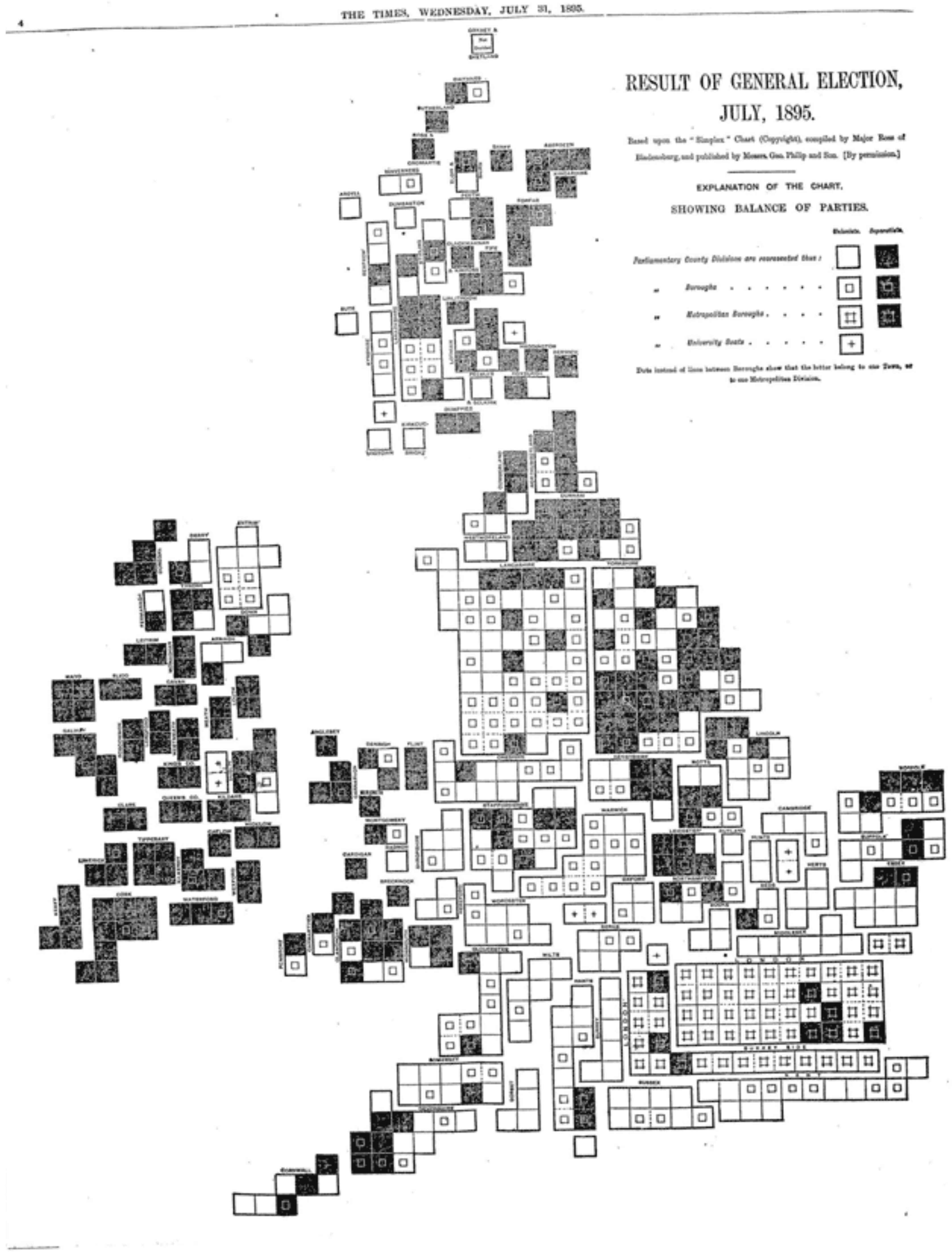
“Everybody says that statistics should be taught. But how?

Statistics are not simply figures. It is said that nothing lies like figures except facts. You want statistics to tell you the truth. You can find truth there if you know how to get at it, and romance, human interest, humor and fascinating revelations as well.”

– Joseph Pulitzer,
The Power of Public Opinion, 1904



timesmachine via Scott Klein


THE TIMES, WEDNESDAY, JULY 31, 1895.
RESULT OF GENERAL ELECTION, JULY, 1895.
Based upon the "Simplex" Chart (Copyright), compiled by Major Ross of Bladensburg, and published by Messrs. Geo. Philip and Son. [By permission.]
EXPLANATION OF THE CHART,
SHOWING BALANCE OF PARTIES.
Unionists.
Separatists.
Parliamentary County Divisions are represented thus:
" Boroughs
" Metropolitan Boroughs
" University Seats
Dots instead of lines between Boroughs show that the latter belong to one Town, or to one Metropolitan Division.

via Scott Klein, above chart

“We started out with this a long time ago—before the Web, before even reasonably simple computers,” says Sarah Cohen, editor of the computer-assisted reporting (CAR) team at The New York Times. As early as the late 1960s, journalists like Philip Meyer and Elliott Jaspin were using social science methods and data analysis—sometimes with the help of mainframe computers—to generate and test their journalistic hypotheses. “That was how a generation of us learned what [computer-assisted reporting] was,” says Cohen.

CAR is a practice that […] for many years existed only at the margins of most newsrooms, the domain of a few motivated reporters.”

–A Brief History of Computer-Assisted Reporting

● Data journalism
Topic

● Computer-assisted reporting
Topic

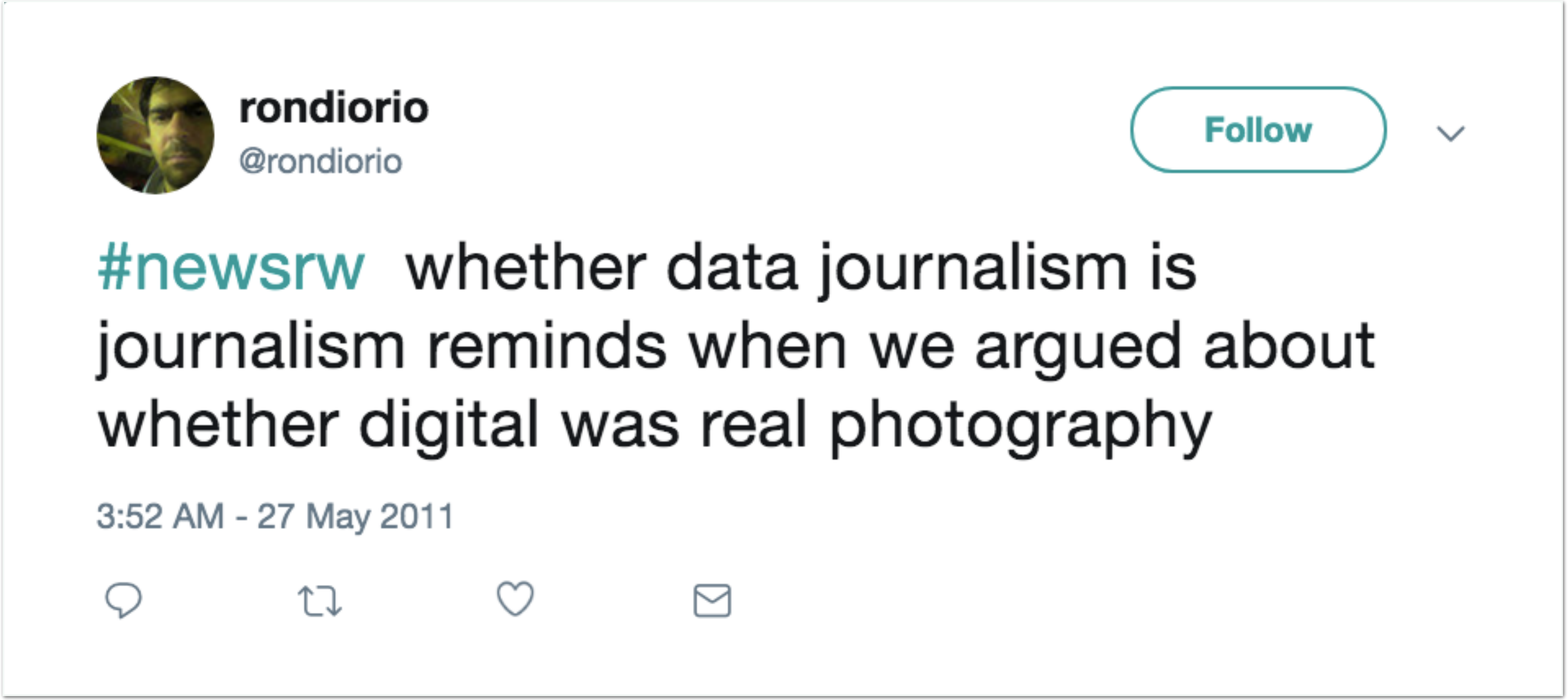
+ Add comparison

United States ▾ 1/1/10 - 6/1/19 ▾ All categories ▾ Web Search ▾

Interest over time

100
75
50
25
Average
Jan 1, 2010
Apr 1, 2013
Jul 1, 2016
Note
Note

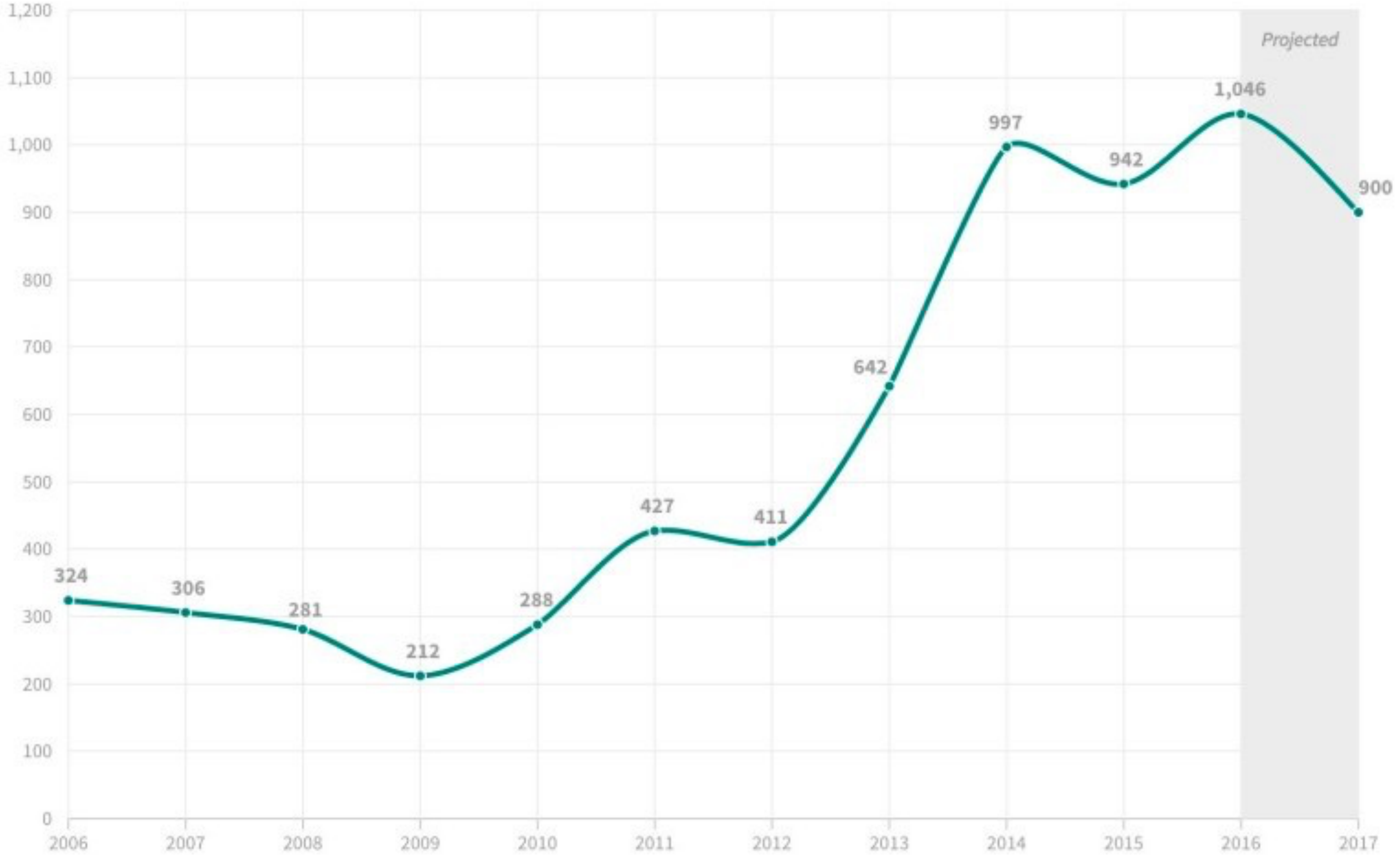
via Google Trends



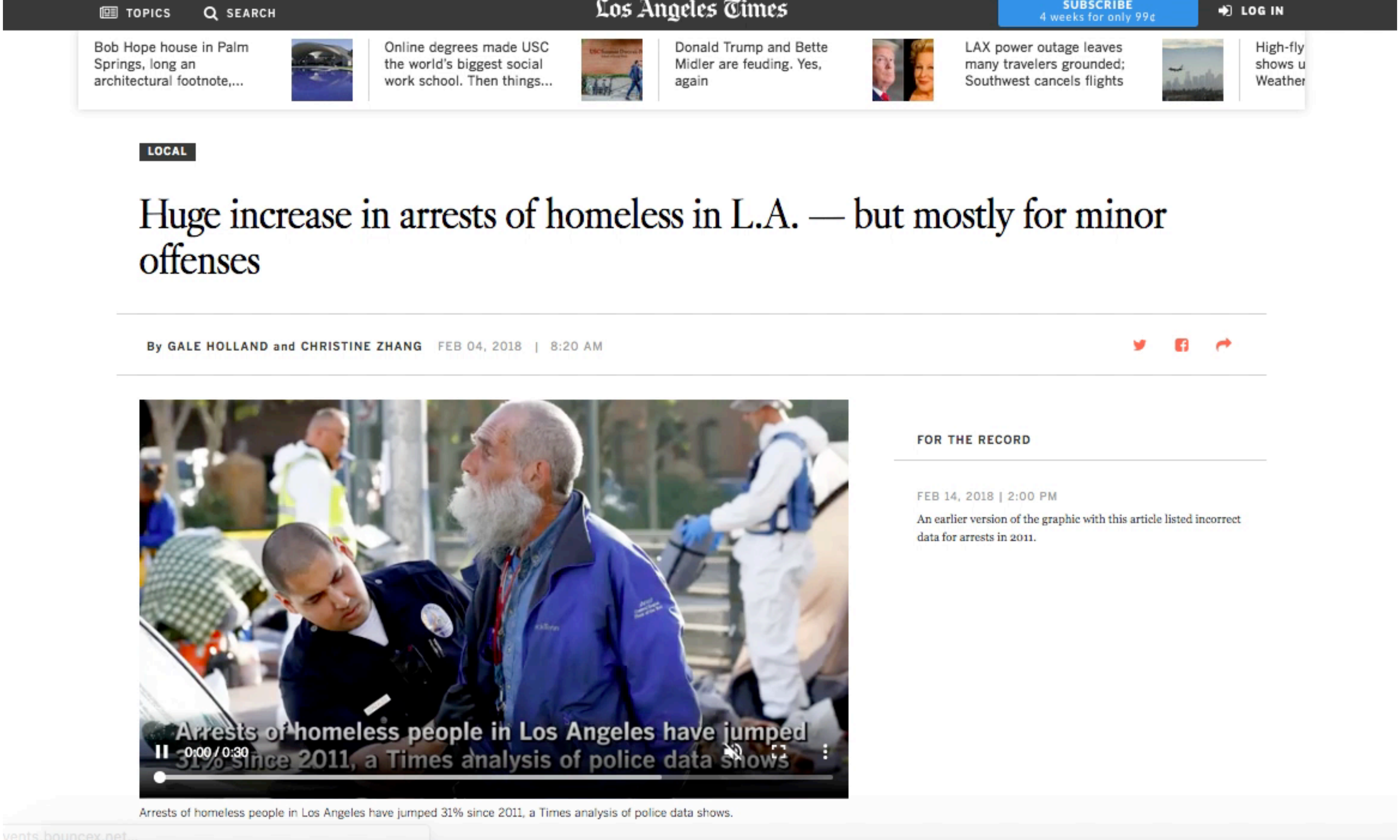
rondiorio
@rondiorio

#newsrw whether data journalism is journalism reminds when we argued about whether digital was real photography

3:52 AM - 27 May 2011



Charting NICAR attendance over the years



TOPICS SEARCH Los Angeles Times SUBSCRIBE 4 weeks for only 99¢ LOG IN

Bob Hope house in Palm Springs, long an architectural footnote,...

Online degrees made USC the world's biggest social work school. Then things...

Donald Trump and Bette Midler are feuding. Yes, again

LAX power outage leaves many travelers grounded; Southwest cancels flights

High-fly shows u Weather

LOCAL

Huge increase in arrests of homeless in L.A. — but mostly for minor offenses

By GALE HOLLAND and CHRISTINE ZHANG FEB 04, 2018 | 8:20 AM

Arrests of homeless people in Los Angeles have jumped 31% since 2011, a Times analysis of police data shows
0:00 / 0:30

Arrests of homeless people in Los Angeles have jumped 31% since 2011, a Times analysis of police data shows.

FOR THE RECORD

FEB 14, 2018 | 2:00 PM

An earlier version of the graphic with this article listed incorrect data for arrests in 2011.

vents.bouncex.net...

Local • Analysis

No, there haven’t been 18 school shootings in 2018. That number is flat wrong.

The horror of Columbine echoes through 19 years of school shooting survivors

Eleven schools since Columbine High School in 1999 have had mass shootings. Accounts by witnesses and survivors are eerily similar. (Video: Monica Akhtar/Photo: Matt McClain/The Washington Post)

By **John Woodrow Cox** and **Steven Rich**

The stunning number swept across the Internet within minutes of the news Wednesday that, yet again, another young man with another semiautomatic rifle had rampaged through a school, this time at Marjory Stoneman Douglas High in South Florida.

The figure originated with Everytown for Gun Safety, a nonprofit group, co-founded by Michael Bloomberg, that works to prevent gun violence and is most famous for its running tally of school shootings.

“This,” the organization tweeted at 4:22 p.m. Wednesday, “is the 18th school shooting in the U.S. in 2018.”

A tweet by Sen. Bernie Sanders (I-Vt.) including the claim had been liked more than 45,000 times by Thursday evening, and one from political analyst Jeff Greenfield had cracked 126,000. New York City Mayor Bill de Blasio tweeted it, too, as did performers Cher and Alexander William and actors Misha Collins and Albert Brooks. News organizations — including MSNBC, ABC News, NBC News, CBS News, Time, MSN, the BBC, the New York Daily News and HuffPost — also used the number in their coverage. By Wednesday night, the top suggested search after typing “18” into Google was “18 school shootings in 2018.”

SECTIONS
HOME
SEARCH
The New York Times
Account
THE UPSHOT
FOLLOW US:
GET THE UPSHOT IN YOUR INBOX
SHARE

Sweden Finds a Simple Way to Improve New Mothers' Health. It Involves Fathers.

Biggest Offender in Outsize Debt: Graduate Schools

THE NEW HEALTH CARE
Does Your Education Level Affect Your Health?

The New Tariffs Against Mexico Signal the U.S. Isn't a Reliable Negotiating Partner

TheUpshot

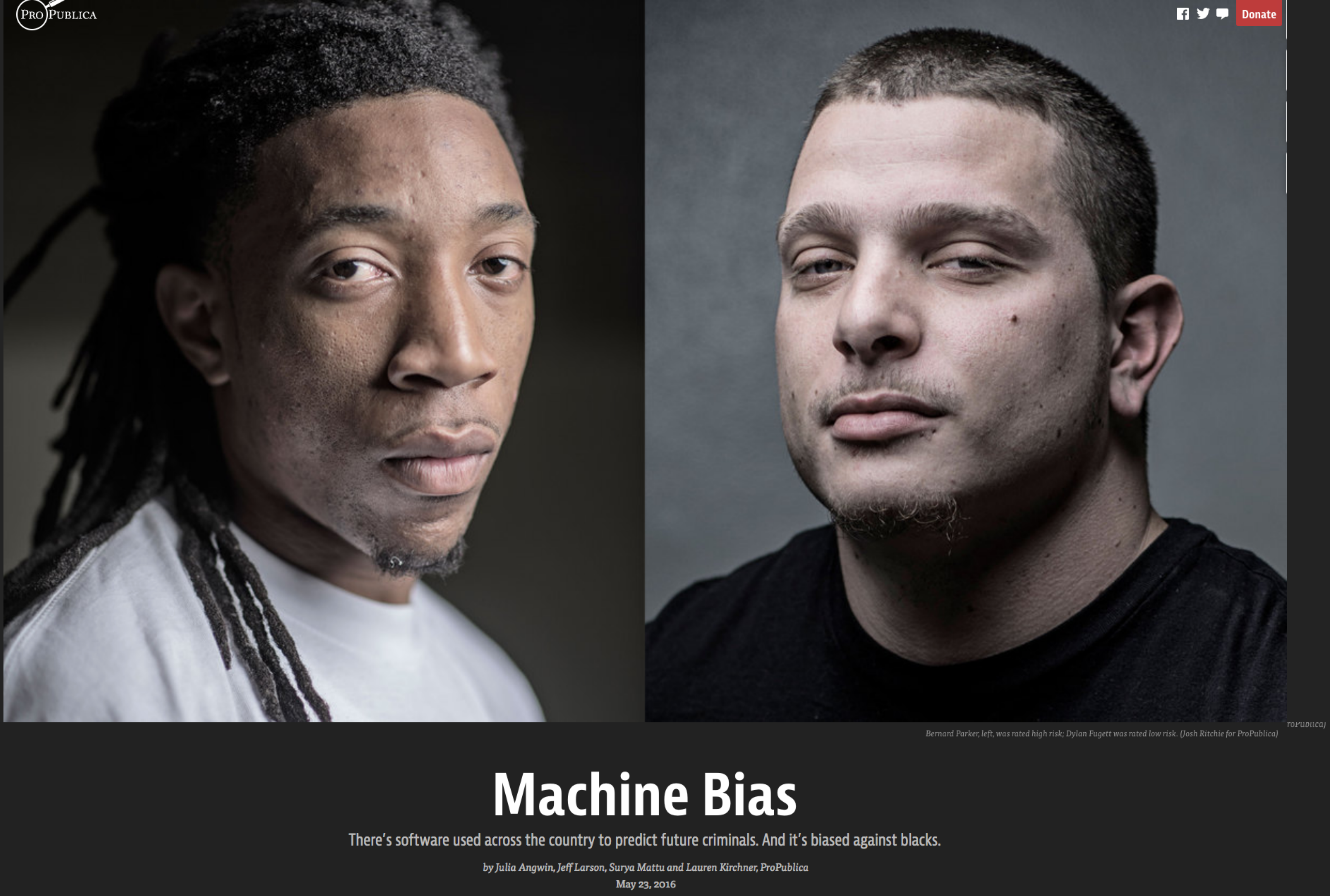
Where the Poor Live Longer: How Your Area Compares

By **GREGOR AISCH, QUOCTRUNG BUI, AMANDA COX** and **KEVIN QUEALY** APRIL 11, 2016

Life expectancy of 40-year-olds with household incomes **below $28,000**, adjusted for race*

Waiting for 5290727.fls.doubleclick.net...

https://www.nytimes.com/interactive/2016/04/11/upshot/where-the-poor-live-longer-how-your-area-compares.html



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

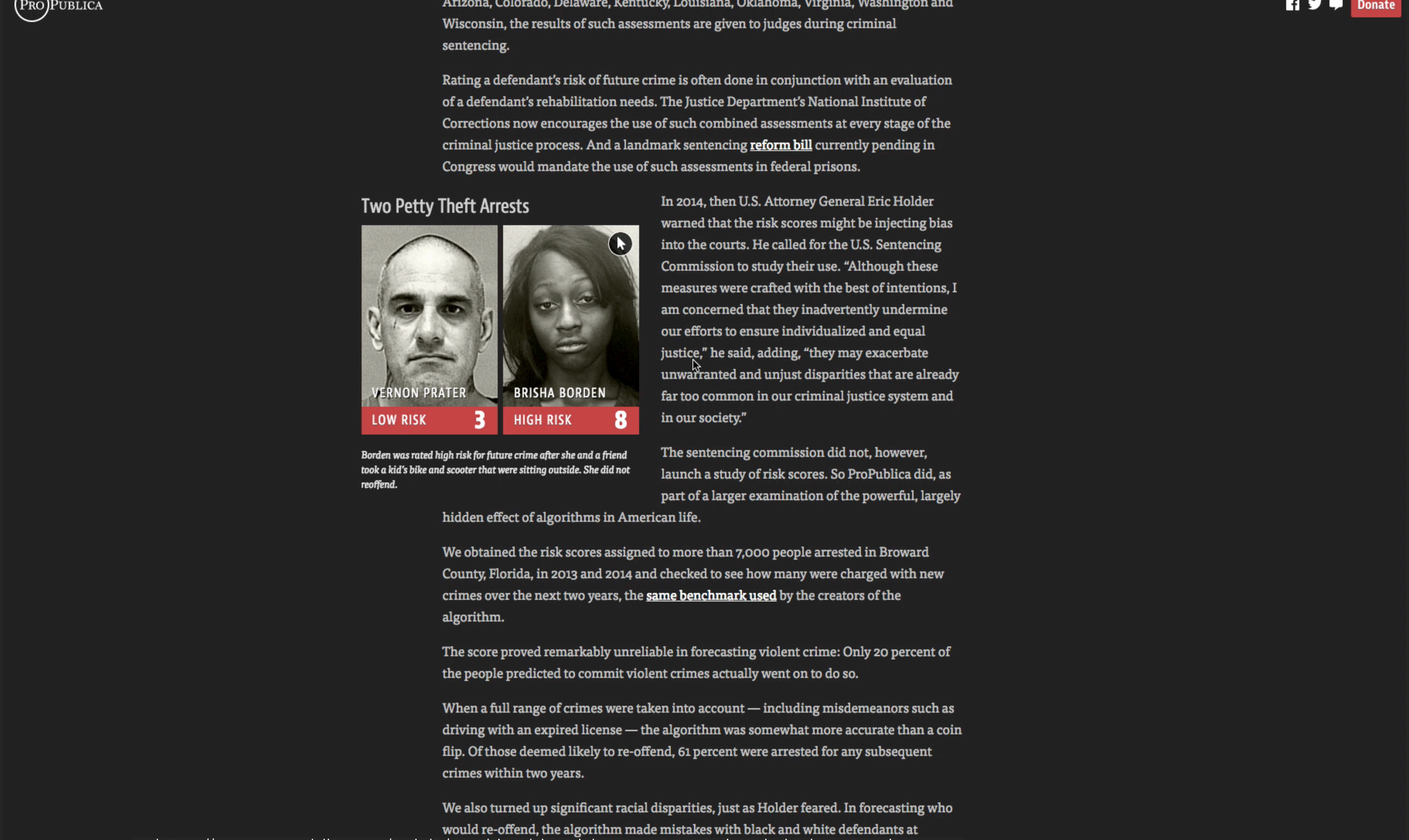
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.

Rating a defendant's risk of future crime is often done in conjunction with an evaluation of a defendant's rehabilitation needs. The Justice Department's National Institute of Corrections now encourages the use of such combined assessments at every stage of the criminal justice process. And a landmark sentencing **reform bill** currently pending in Congress would mandate the use of such assessments in federal prisons.

Two Petty Theft Arrests

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing Commission to study their use. "Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice," he said, adding, "they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."

The sentencing commission did not, however, launch a study of risk scores. So ProPublica did, as part of a larger examination of the powerful, largely hidden effect of algorithms in American life.

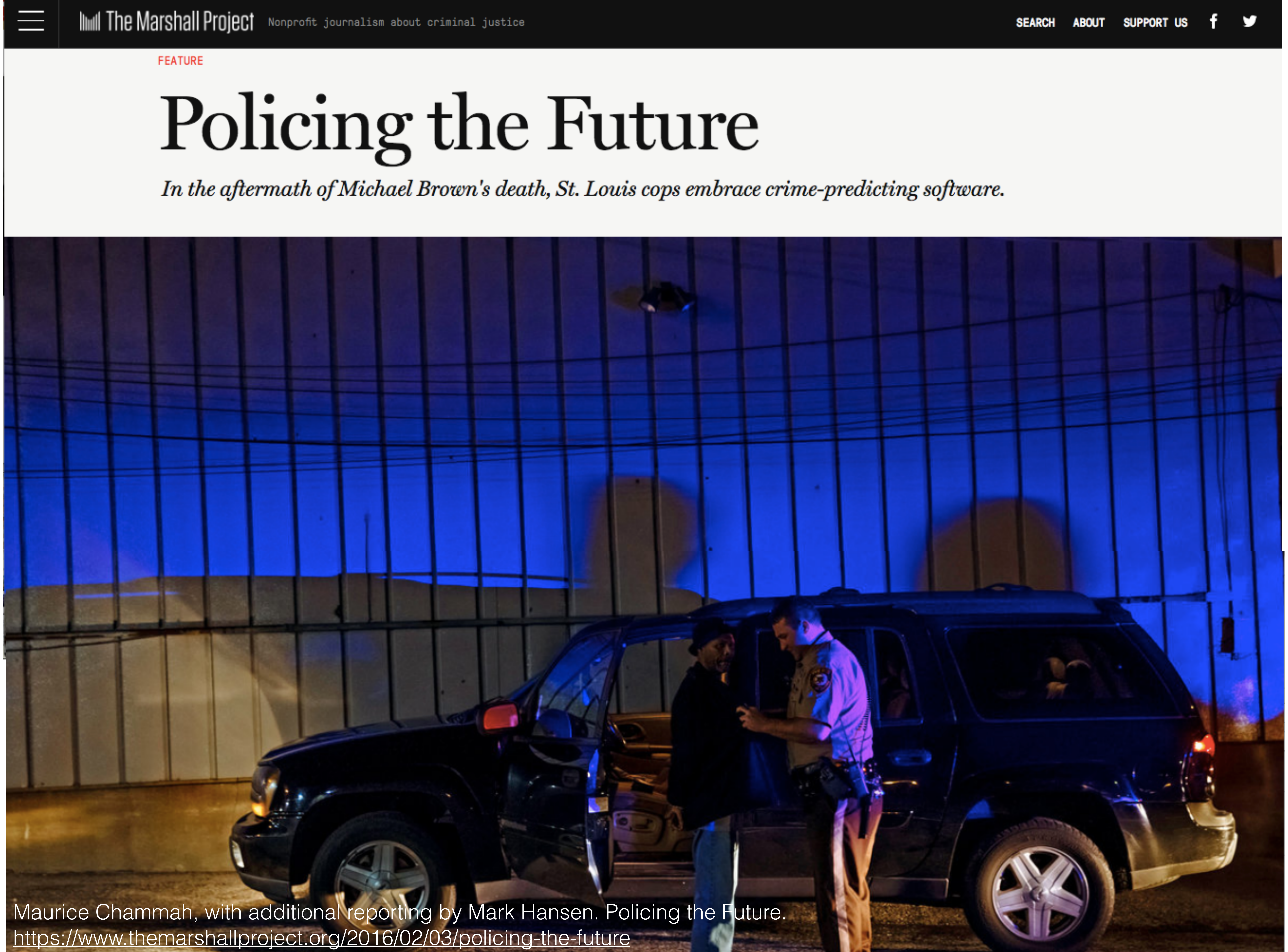
We obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the **same benchmark used** by the creators of the algorithm.

The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.

When a full range of crimes were taken into account — including misdemeanors such as driving with an expired license — the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.

We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

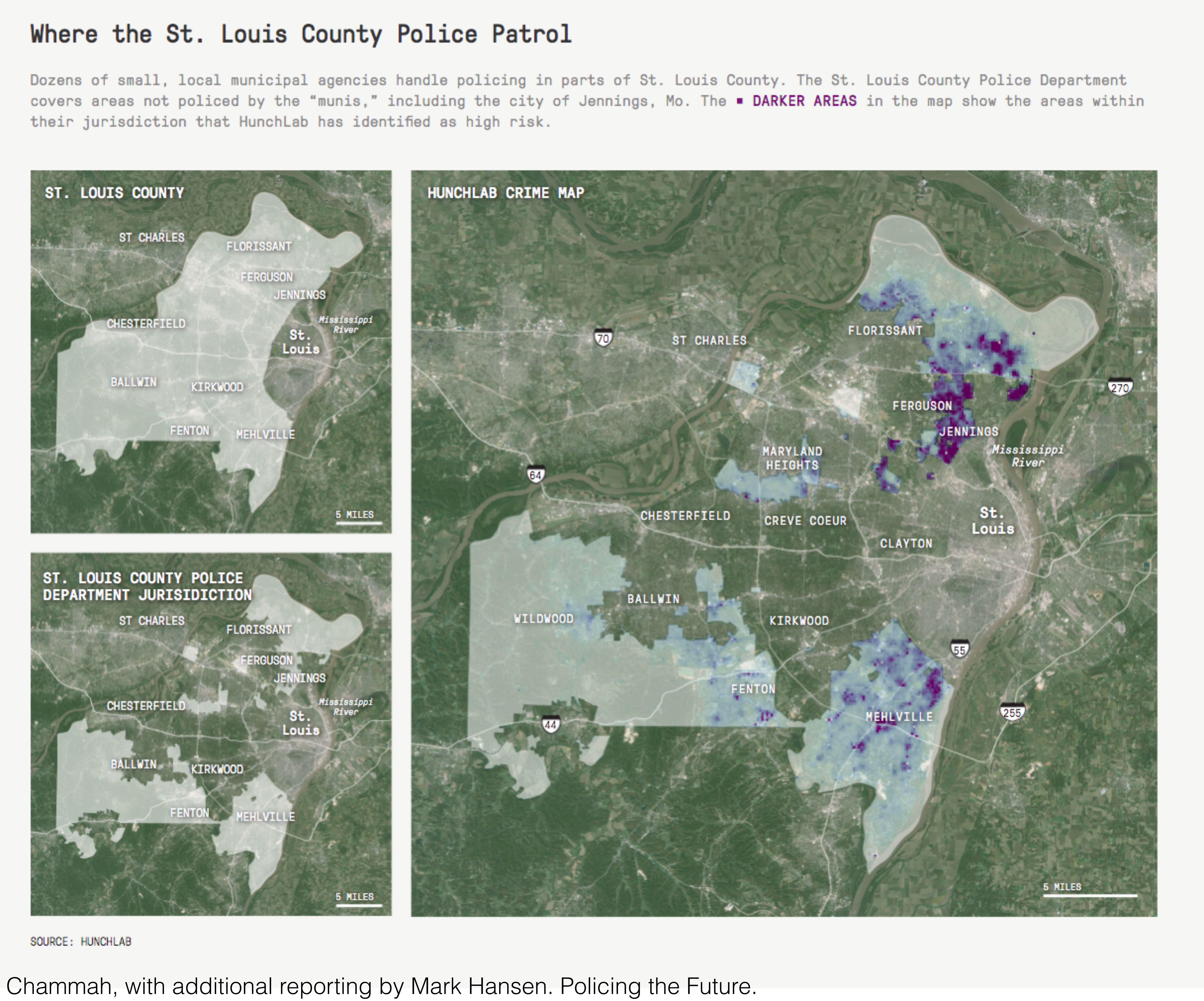


FEATURE

Policing the Future

In the aftermath of Michael Brown's death, St. Louis cops embrace crime-predicting software.

Maurice Chammah, with additional reporting by Mark Hansen. Policing the Future.
https://www.themarshallproject.org/2016/02/03/policing-the-future

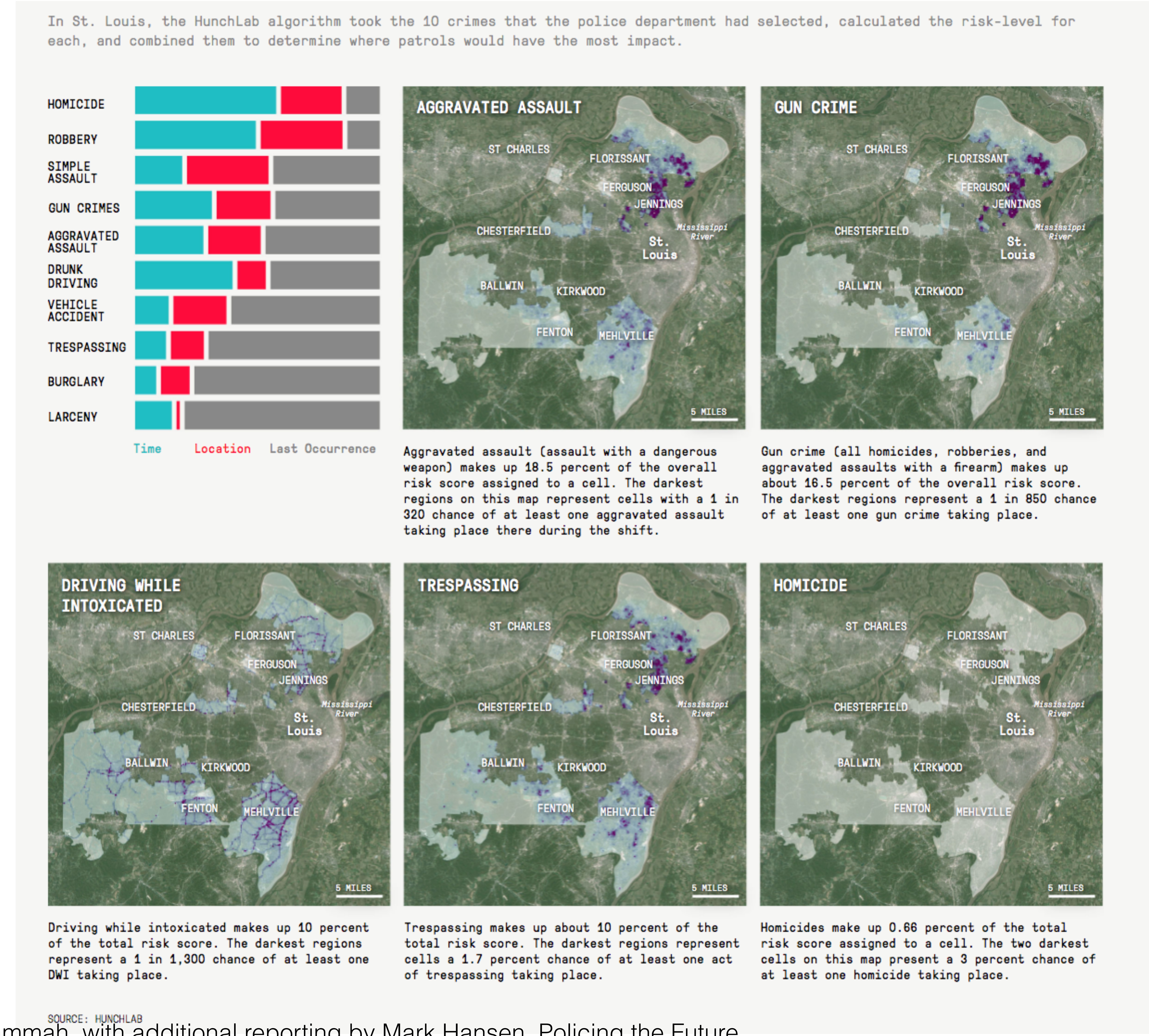


Where the St. Louis County Police Patrol

Dozens of small, local municipal agencies handle policing in parts of St. Louis County. The St. Louis County Police Department covers areas not policed by the “munis,” including the city of Jennings, Mo. The ■ DARKER AREAS in the map show the areas within their jurisdiction that HunchLab has identified as high risk.

SOURCE: HUNCHLAB

Maurice Chammah, with additional reporting by Mark Hansen. Policing the Future.
https://www.themarshallproject.org/2016/02/03/policing-the-future



In St. Louis, the HunchLab algorithm took the 10 crimes that the police department had selected, calculated the risk-level for each, and combined them to determine where patrols would have the most impact.

AGGRAVATED ASSAULT

Aggravated assault (assault with a dangerous weapon) makes up 18.5 percent of the overall risk score assigned to a cell. The darkest regions on this map represent cells with a 1 in 320 chance of at least one aggravated assault taking place there during the shift.

GUN CRIME

Gun crime (all homicides, robberies, and aggravated assaults with a firearm) makes up about 16.5 percent of the overall risk score. The darkest regions represent a 1 in 850 chance of at least one gun crime taking place.

DRIVING WHILE INTOXICATED

Driving while intoxicated makes up 10 percent of the total risk score. The darkest regions represent a 1 in 1,300 chance of at least one DWI taking place.

TRESPASSING

Trespassing makes up about 10 percent of the total risk score. The darkest regions represent cells a 1.7 percent chance of at least one act of trespassing taking place.

HOMICIDE

Homicides make up 0.66 percent of the total risk score assigned to a cell. The two darkest cells on this map present a 3 percent chance of at least one homicide taking place.

SOURCE: HUNCHLAB

Maurice Chammah, with additional reporting by Mark Hansen. Policing the Future. https://www.themarshallproject.org/2016/02/03/policing-the-future

What is data journalism?

flickr: jdhancock

Definitely DJ
dataset
has a data visualization
- "new study says" (may or may not be newsworthy)
not all studies are quantitative
Election reporting
Story is about a dataset you collected or an experiment you did
Could be DJ
includes #s: poll numbers, but it's not central to the story
comparison of numbers this tweet got many times more RTs
opinionated criticism
use of data by companies
Weather??
Not DJ
current events, w/o #s or analysis
This tweet got 5000 RTs
just about a person
Opinion pieces
- needs comparison or context
- depends on perspective
- journalistic
- reader

Definitely data journalism

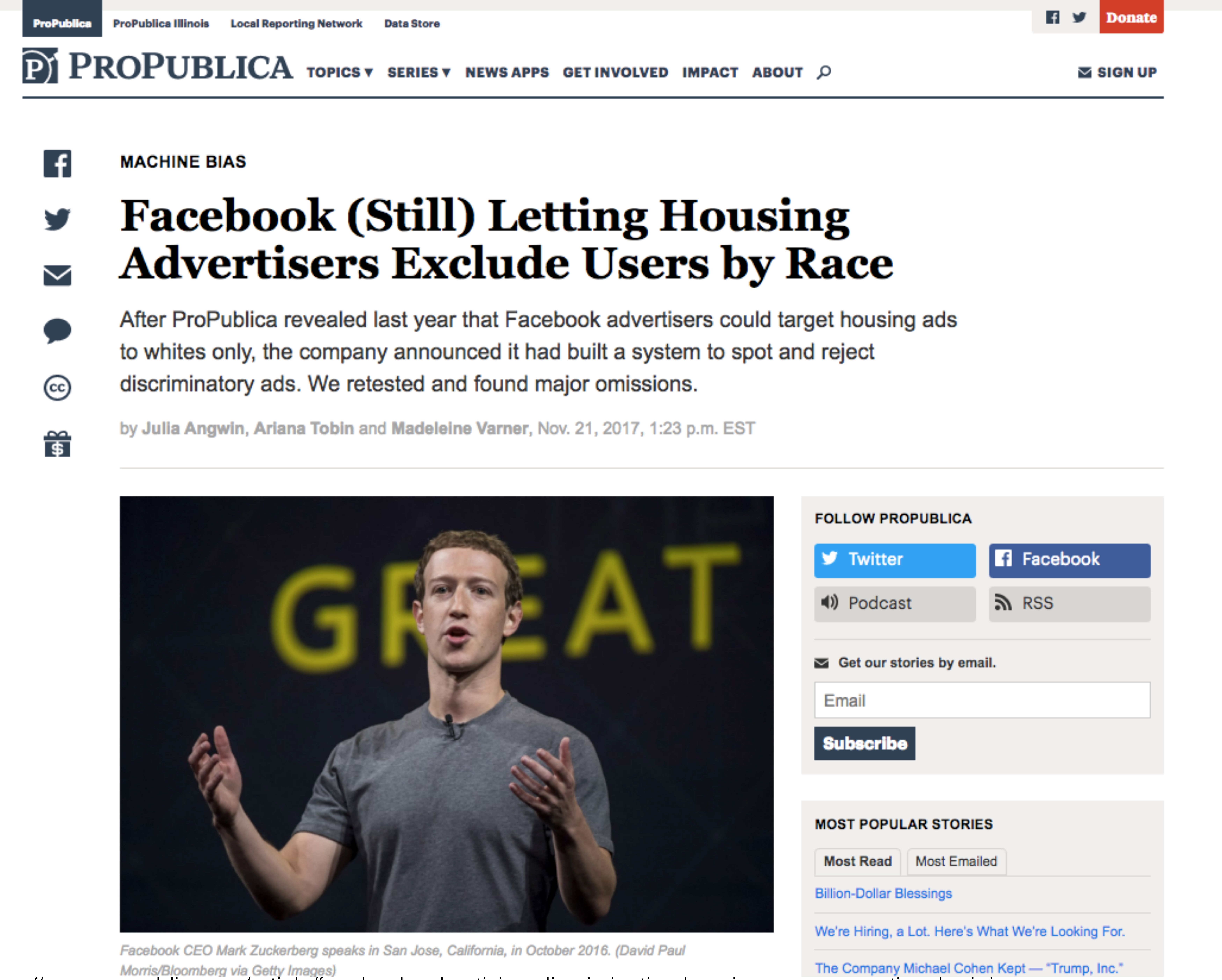
- the story is about the examination of a dataset
- has a data visualization
- "new study says" (may or may not be newsworthy)
- election reporting
- story is about a dataset you collected or an experiment you did

Could be data journalism

- includes numbers (e.g. poll numbers) but not central to the story
- comparison of numbers (e.g. "this tweet got [x] times as many RTs")
- opinionated criticism of a dataset or study
- about the use of data by companies
- weather??

Definitely not data journalism

- current events, without numbers or analysis
- profile of a person
- opinion pieces



MACHINE BIAS

Facebook (Still) Letting Housing Advertisers Exclude Users by Race

After ProPublica revealed last year that Facebook advertisers could target housing ads to whites only, the company announced it had built a system to spot and reject discriminatory ads. We retested and found major omissions.

by Julia Angwin, Ariana Tobin and Madeleine Varner, Nov. 21, 2017, 1:23 p.m. EST

Facebook CEO Mark Zuckerberg speaks in San Jose, California, in October 2016. (David Paul Morris/Bloomberg via Getty Images)

Billion-Dollar Blessings

We're Hiring, a Lot. Here's What We're Looking For.

The Company Michael Cohen Kept — "Trump, Inc."

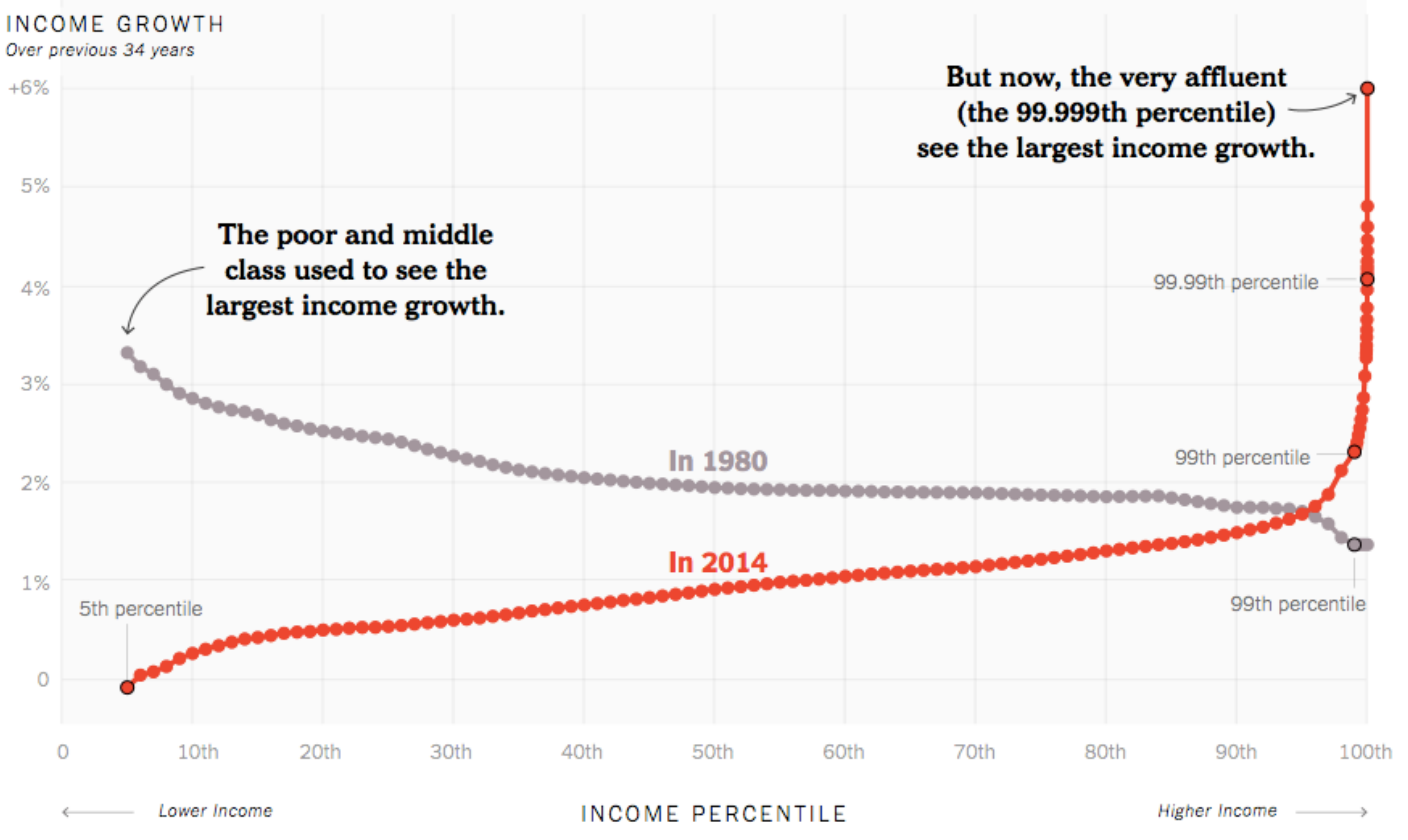
https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin

Opinion

Our Broken Economy, in One Simple Chart

By DAVID LEONHARDT

AUG. 7, 2017



Note: Inflation-adjusted annual average growth using income after taxes, transfers and non-cash benefits.

flickr: jdhancock

A course in data journalism

SDS 236 - Spring 2018
Syllabus
Schedule
Assignments
Resources

Schedule

Week 1

This is a short week, with only one class. We'll mostly be getting to know one another and getting up to speed with Google Drive and Slack. By Sunday night, you need to have introduced yourself on Slack and written your first Data Diary entry. For Tuesday in class, you need to come prepared with some links for your Wikipedia article.

1/25

- Introduction to the practice of data journalism
- Paper and pencil data collection on NYTimes
- What is newsworthy?

Week 2

On Tuesday, we wrapped up our NYTimes data collection, discussed chapter 1 of Numbers in the Newsroom, and talked through the wikipedia authoring process. Thursday class was cancelled.

Wikipedia entries are due Monday at midnight.

By Tuesday, I'd like you to have read Numbers in the Newsroom through the section "Going further with changes" (where through means you read that section but you don't need to go beyond), and Data Organization in Spreadsheets. We'll be starting to interview a spreadsheet during our Zoom meeting.

1/30

- Wrapping up NYTimes data collection
- Discussing Chapter 1 of Numbers in the Newsroom

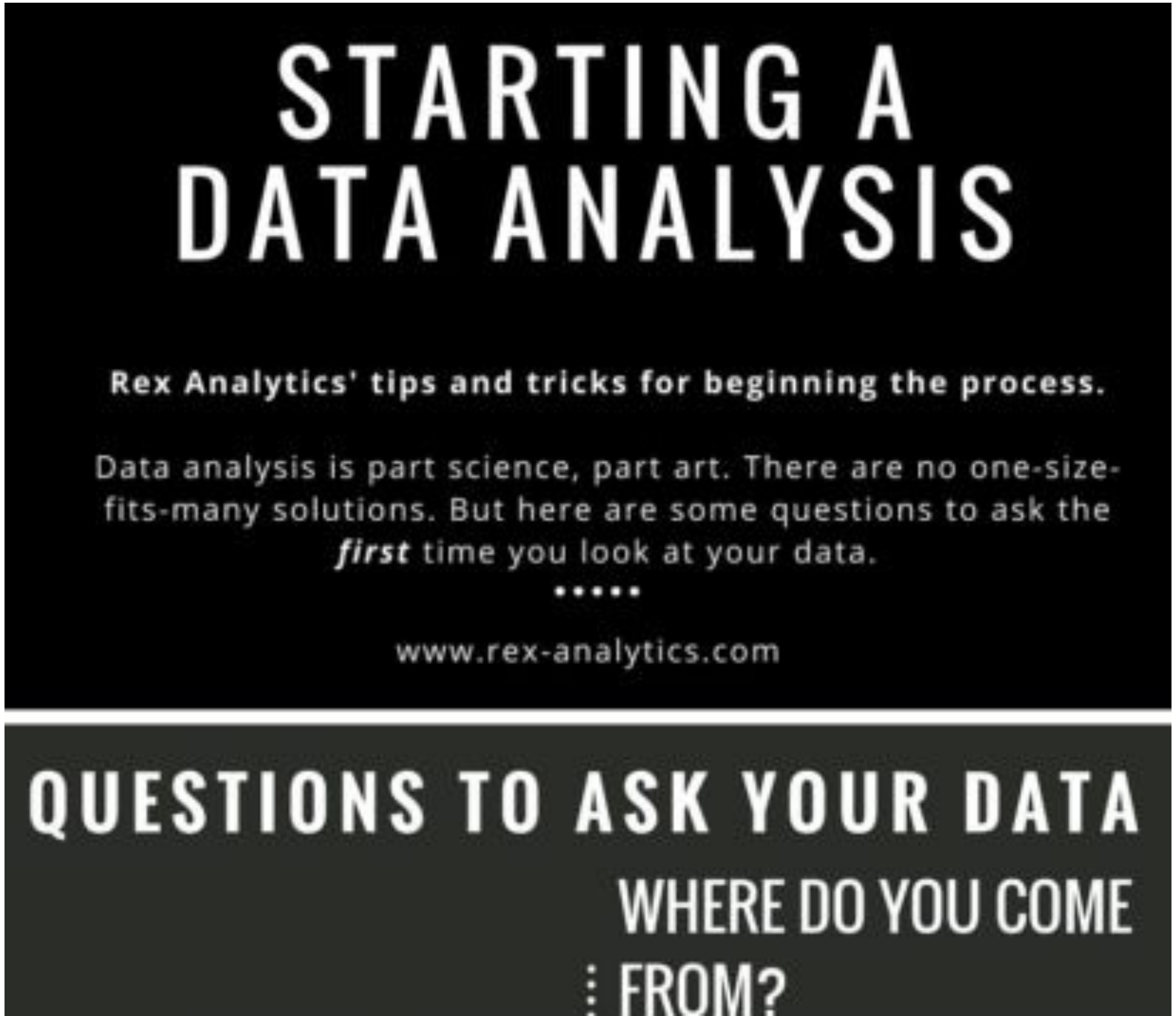
2/1

- No class

Interviewing

- Know your subject
- Come in with a plan
- Write questions ahead of time, but prioritize conversation
- Just come out and ask the hard stuff
- Embrace the silences
- Think in soundbites
- Play dumb
- Oh, and finally, “Keep the mic running after you finish”

The Art of the Interview



STARTING A DATA ANALYSIS

Rex Analytics' tips and tricks for beginning the process.

Data analysis is part science, part art. There are no one-size-fits-many solutions. But here are some questions to ask the ***first*** time you look at your data.

•••••

www.rex-analytics.com

QUESTIONS TO ASK YOUR DATA

WHERE DO YOU COME FROM?

🔦 One-number story

"Keep the number of digits in a paragraph below eight."

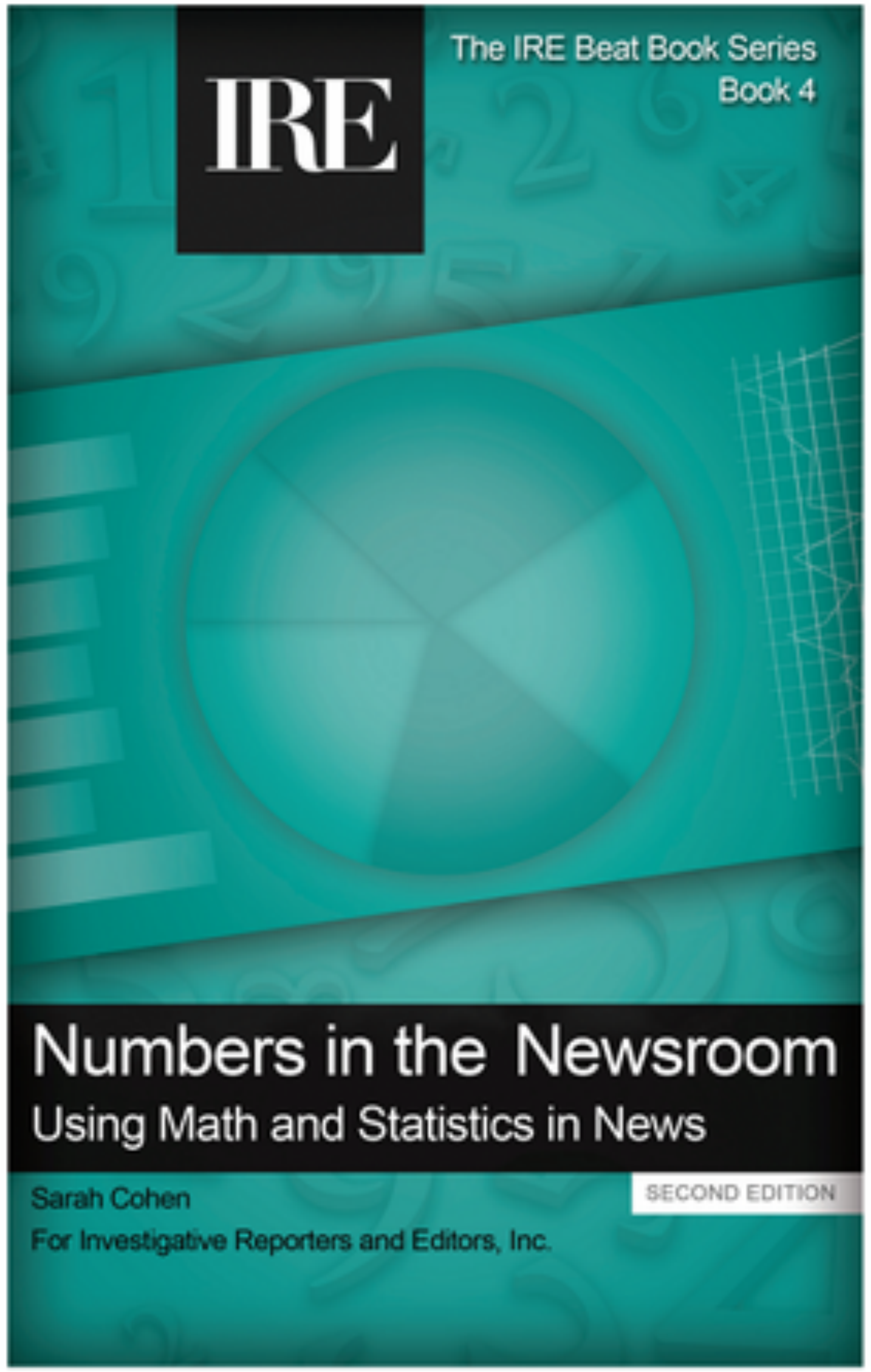
"You'd be over your allocation with a sentence like this:

The Office of Redundancy's budget rose 48 percent in 2013, from $700.3 million to $1.03 billion.

Think about how it could change:

Over the past year, the Office of Redundancy's budget grew by nearly half, to $1 billion."

- Sarah Cohen, Numbers in the Newsroom

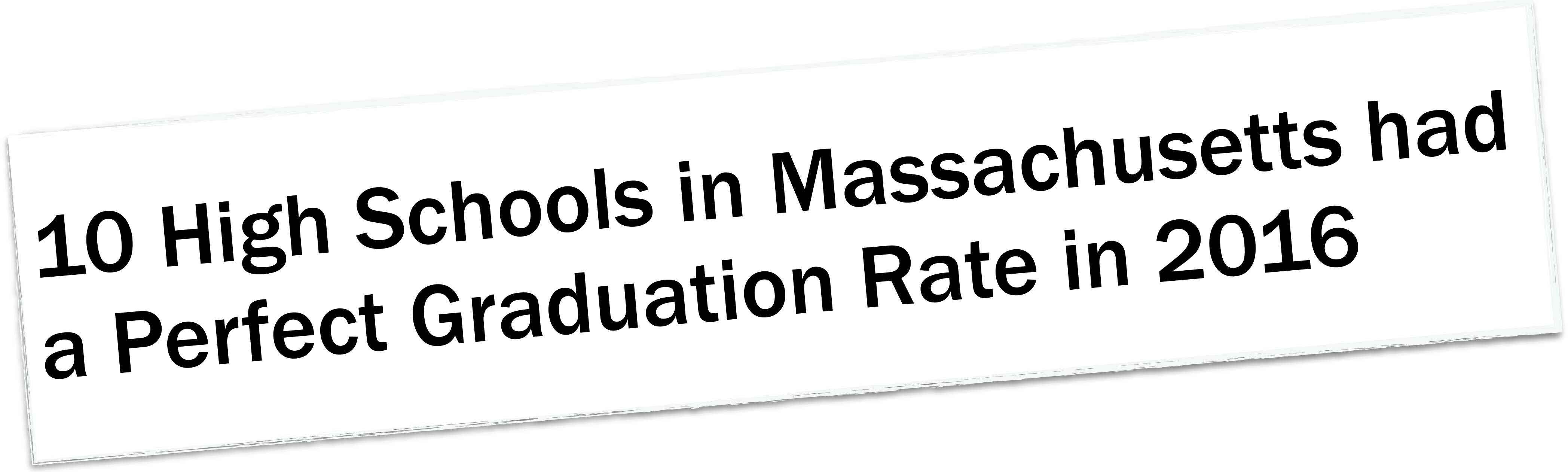


🔦 One-number story

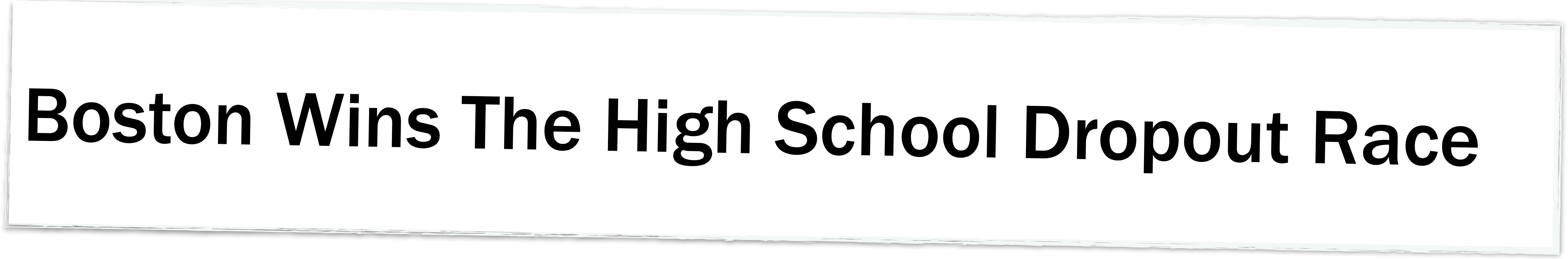
Focus on one number (but use more numbers to contextualize it!)

That number might be the mean, the median, the maximum, the total…

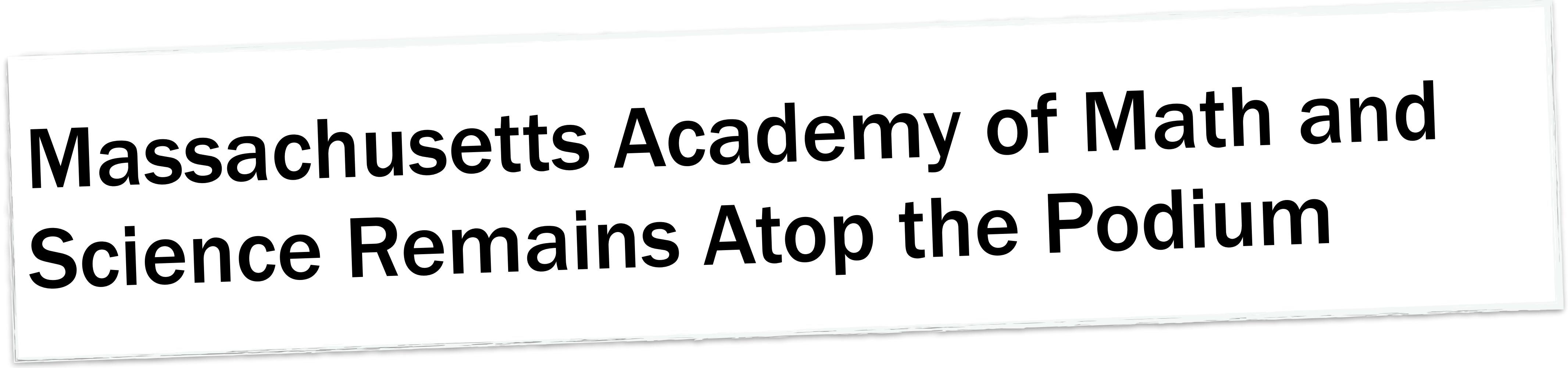
Use simple data tools— in my class, we use spreadsheets for this assignment (sort, summarize, pivot tables)



10 High Schools in Massachusetts had a Perfect Graduation Rate in 2016

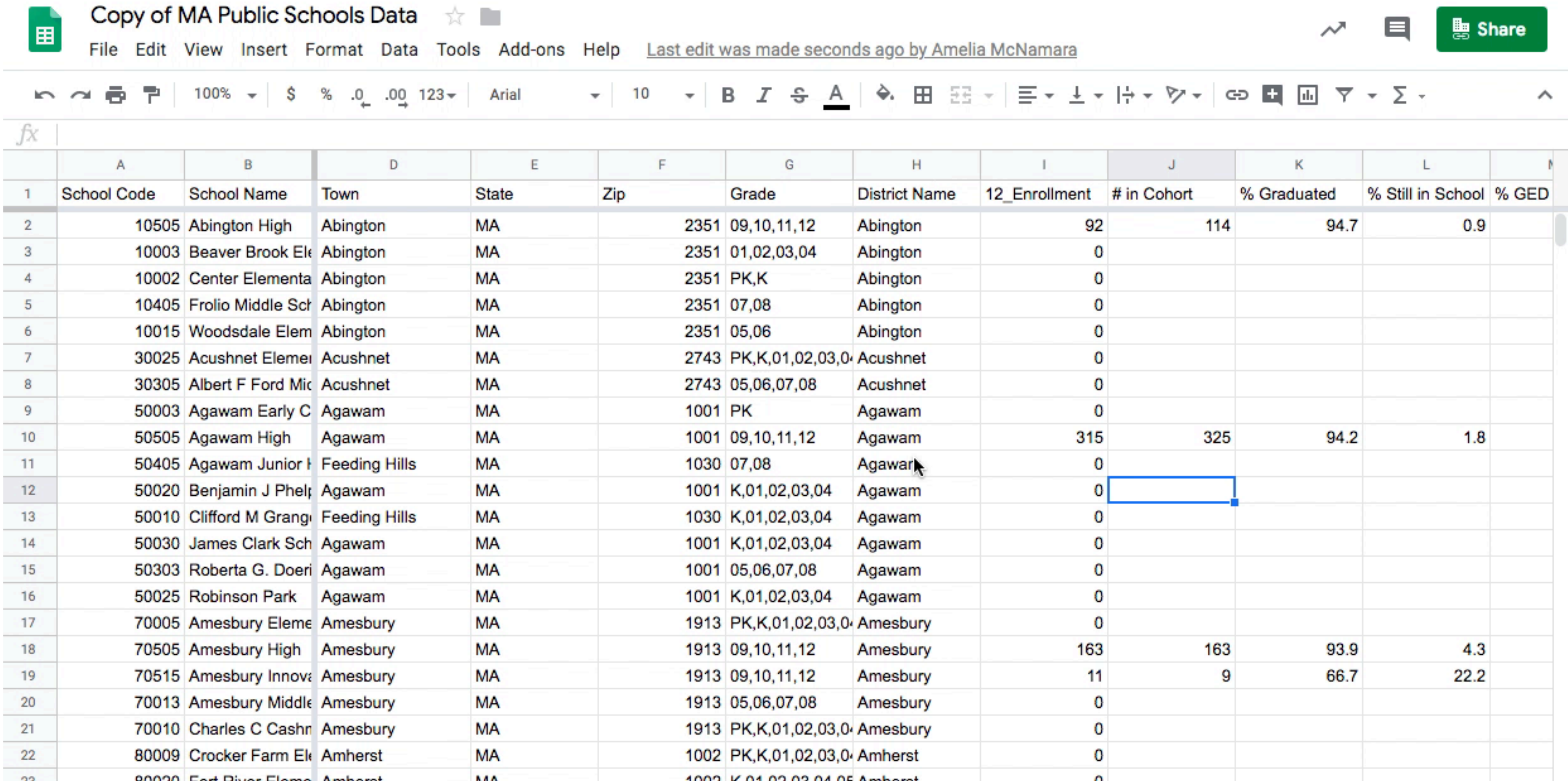


Boston Wins The High School Dropout Race



Massachusetts Academy of Math and Science Remains Atop the Podium

🔦 One-number story

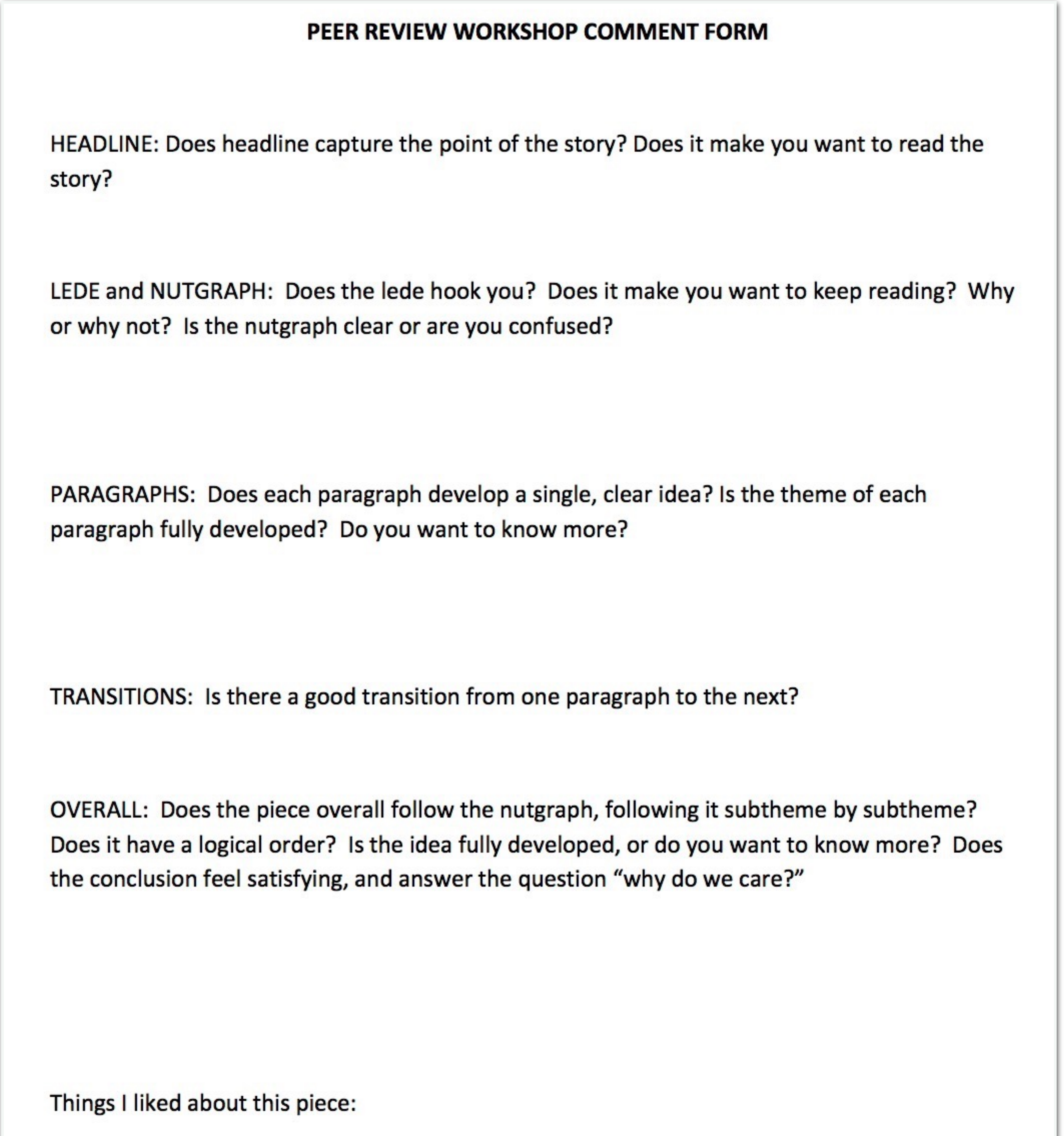

Copy of MA Public Schools Data
File
Edit
View
Insert
Format
Data
Tools
Add-ons
Help
Last edit was made seconds ago by Amelia McNamara
Share
100%
$
%
.0
.00
123
Arial
10
B
I
S
A
fx

	A	B	D	E	F	G	H	I	J	K	L	M
1	School Code	School Name	Town	State	Zip	Grade	District Name	12_Enrollment	# in Cohort	% Graduated	% Still in School	% GED
2	10505	Abington High	Abington	MA	2351	09,10,11,12	Abington	92	114	94.7	0.9	
3	10003	Beaver Brook Ele	Abington	MA	2351	01,02,03,04	Abington	0				
4	10002	Center Elementa	Abington	MA	2351	PK,K	Abington	0				
5	10405	Frolio Middle Sch	Abington	MA	2351	07,08	Abington	0				
6	10015	Woodsdale Elem	Abington	MA	2351	05,06	Abington	0				
7	30025	Acushnet Elemer	Acushnet	MA	2743	PK,K,01,02,03,0	Acushnet	0				
8	30305	Albert F Ford Mic	Acushnet	MA	2743	05,06,07,08	Acushnet	0				
9	50003	Agawam Early C	Agawam	MA	1001	PK	Agawam	0				
10	50505	Agawam High	Agawam	MA	1001	09,10,11,12	Agawam	315	325	94.2	1.8	
11	50405	Agawam Junior H	Feeding Hills	MA	1030	07,08	Agawam	0				
12	50020	Benjamin J Phelp	Agawam	MA	1001	K,01,02,03,04	Agawam	0				
13	50010	Clifford M Grang	Feeding Hills	MA	1030	K,01,02,03,04	Agawam	0				
14	50030	James Clark Sch	Agawam	MA	1001	K,01,02,03,04	Agawam	0				
15	50303	Roberta G. Doeri	Agawam	MA	1001	05,06,07,08	Agawam	0				
16	50025	Robinson Park	Agawam	MA	1001	K,01,02,03,04	Agawam	0				
17	70005	Amesbury Eleme	Amesbury	MA	1913	PK,K,01,02,03,0	Amesbury	0				
18	70505	Amesbury High	Amesbury	MA	1913	09,10,11,12	Amesbury	163	163	93.9	4.3	
19	70515	Amesbury Innova	Amesbury	MA	1913	09,10,11,12	Amesbury	11	9	66.7	22.2	
20	70013	Amesbury Middle	Amesbury	MA	1913	05,06,07,08	Amesbury	0				
21	70010	Charles C Cashn	Amesbury	MA	1913	PK,K,01,02,03,0	Amesbury	0				
22	80009	Crocker Farm Ele	Amherst	MA	1002	PK,K,01,02,03,0	Amherst	0				
23	80020	Fort River Eleme	Amherst	MA	1002	K,01,02,03,04,05	Amherst	0				

🔦 One-number story

Again, iteration is key

- First draft ➡️
- Peer editing in class ➡️
- Final draft ➡️
- Feedback from professor



PEER REVIEW WORKSHOP COMMENT FORM

HEADLINE: Does headline capture the point of the story? Does it make you want to read the story?

LEDE and NUTGRAPH: Does the lede hook you? Does it make you want to keep reading? Why or why not? Is the nutgraph clear or are you confused?

PARAGRAPHS: Does each paragraph develop a single, clear idea? Is the theme of each paragraph fully developed? Do you want to know more?

TRANSITIONS: Is there a good transition from one paragraph to the next?

OVERALL: Does the piece overall follow the nutgraph, following it subtheme by subtheme? Does it have a logical order? Is the idea fully developed, or do you want to know more? Does the conclusion feel satisfying, and answer the question "why do we care?"

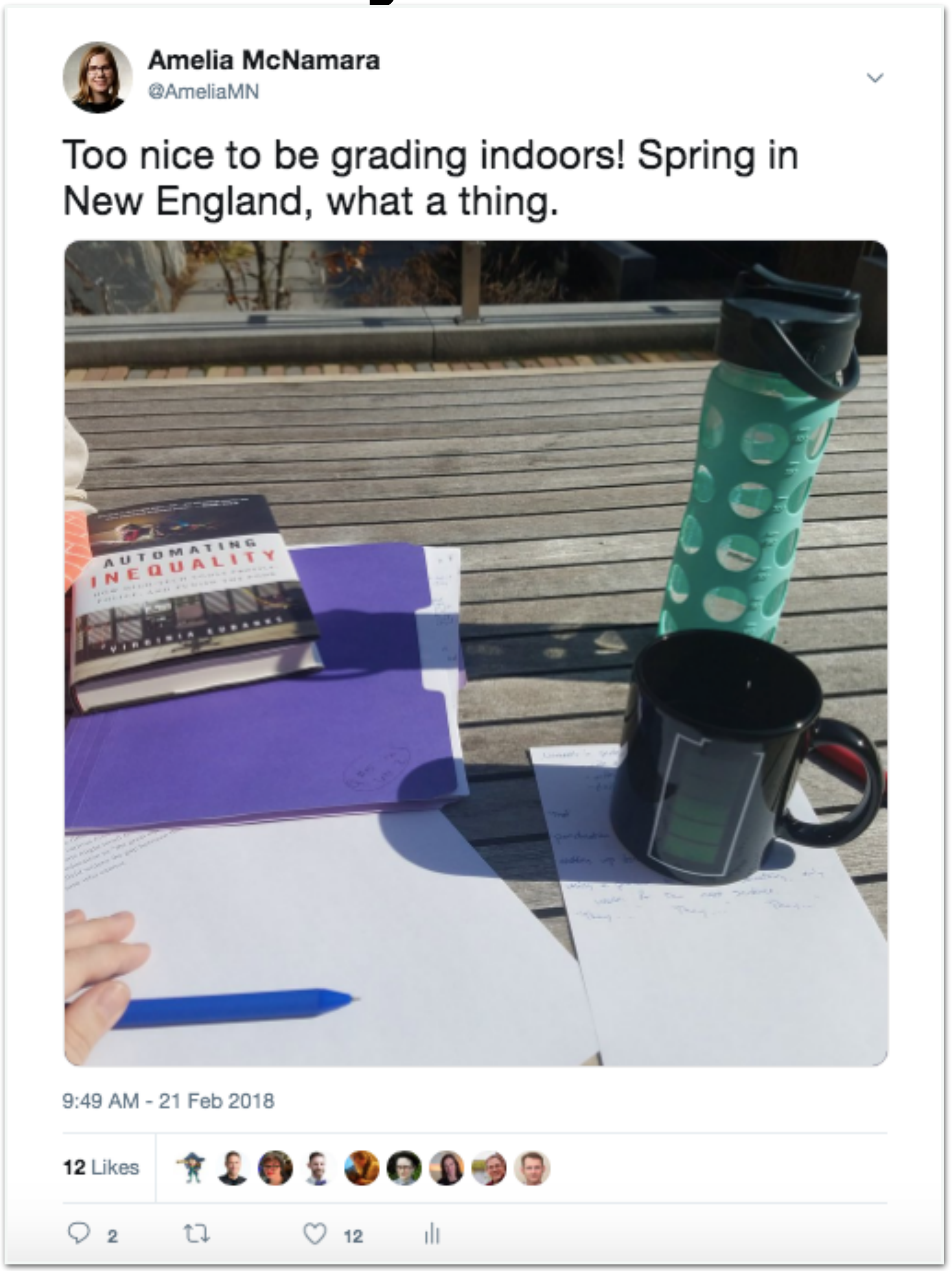
Things I liked about this piece:

🔦 One-number story

Have students turn in hard copies, or print them out

Usually when we grade writing we ignore or give little weight to things like grammar, sentence structure, and awkward paragraphs.

This is the place to give feedback on those things.



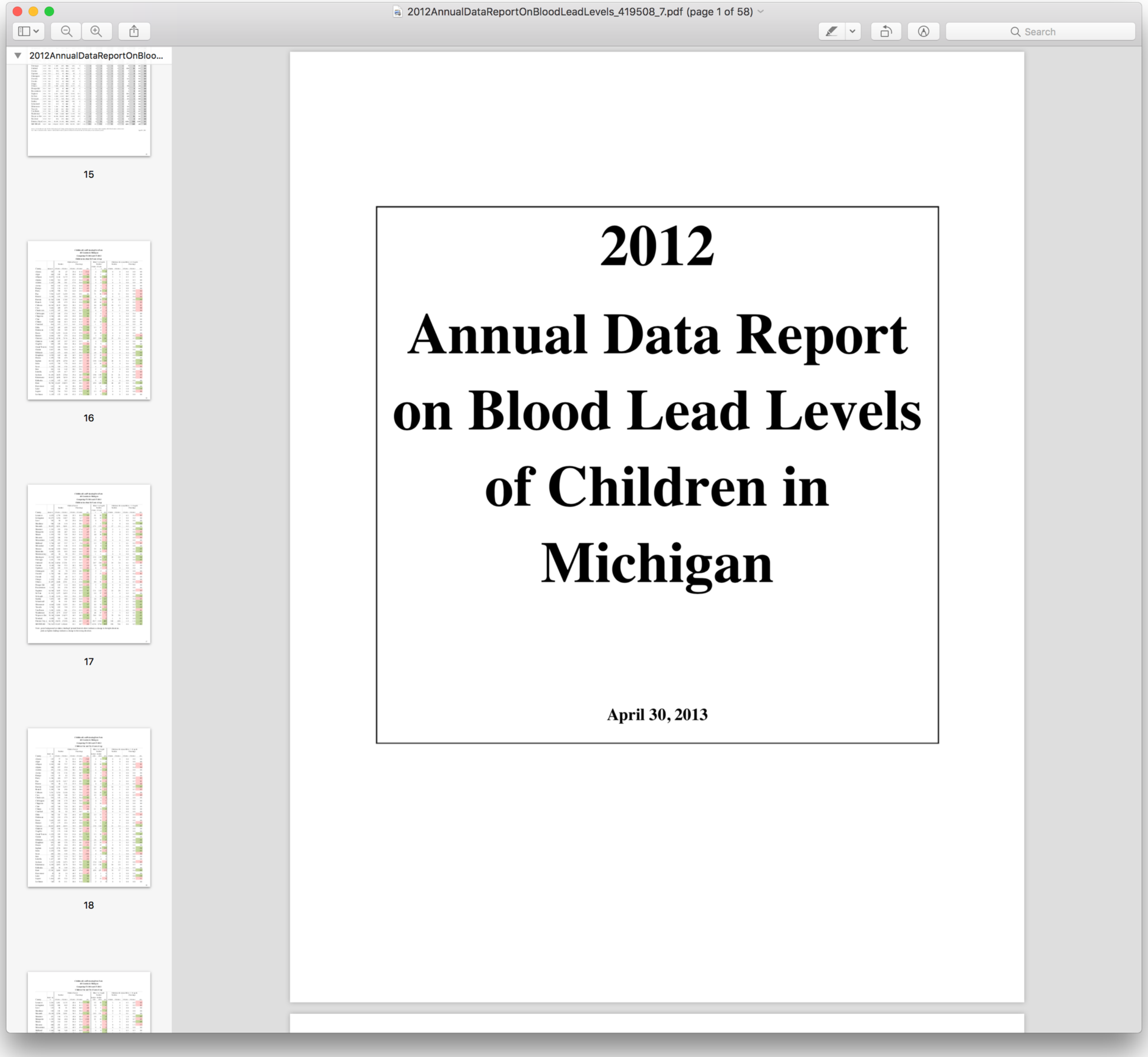
Amelia McNamara
@AmeliaMN

Too nice to be grading indoors! Spring in New England, what a thing.

9:49 AM - 21 Feb 2018

12 Likes

🔦 Freeing data from PDFS using Tabula



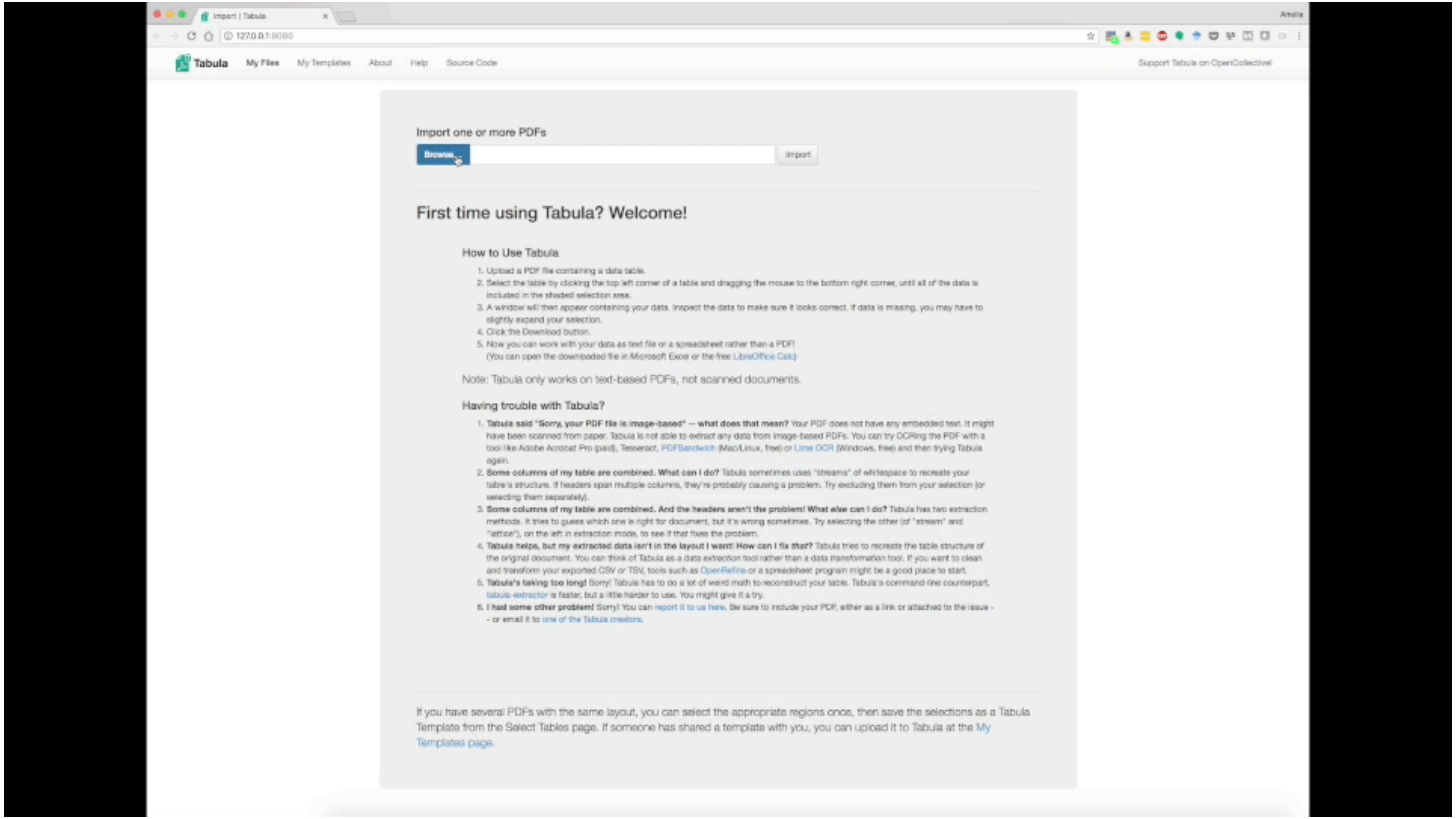
2012AnnualDataReportOnBloodLeadLevels_419508_7.pdf (page 1 of 58)

2012

Annual Data Report on Blood Lead Levels of Children in Michigan

April 30, 2013

🔦 Freeing data from PDFs using Tabula



First time using Tabula? Welcome!

How to Use Tabula

1. Upload a PDF file containing a data table.
2. Select the table by clicking the top left corner of a table and dragging the mouse to the bottom right corner, until all of the data is included in the shaded selection area.
3. A window will then appear containing your data. Inspect the data to make sure it looks correct. If data is missing, you may have to slightly expand your selection.
4. Click the Download button.
5. Now you can work with your data as text file or a spreadsheet rather than a PDF!
   (You can open the downloaded file in Microsoft Excel or the free LibreOffice Calc)

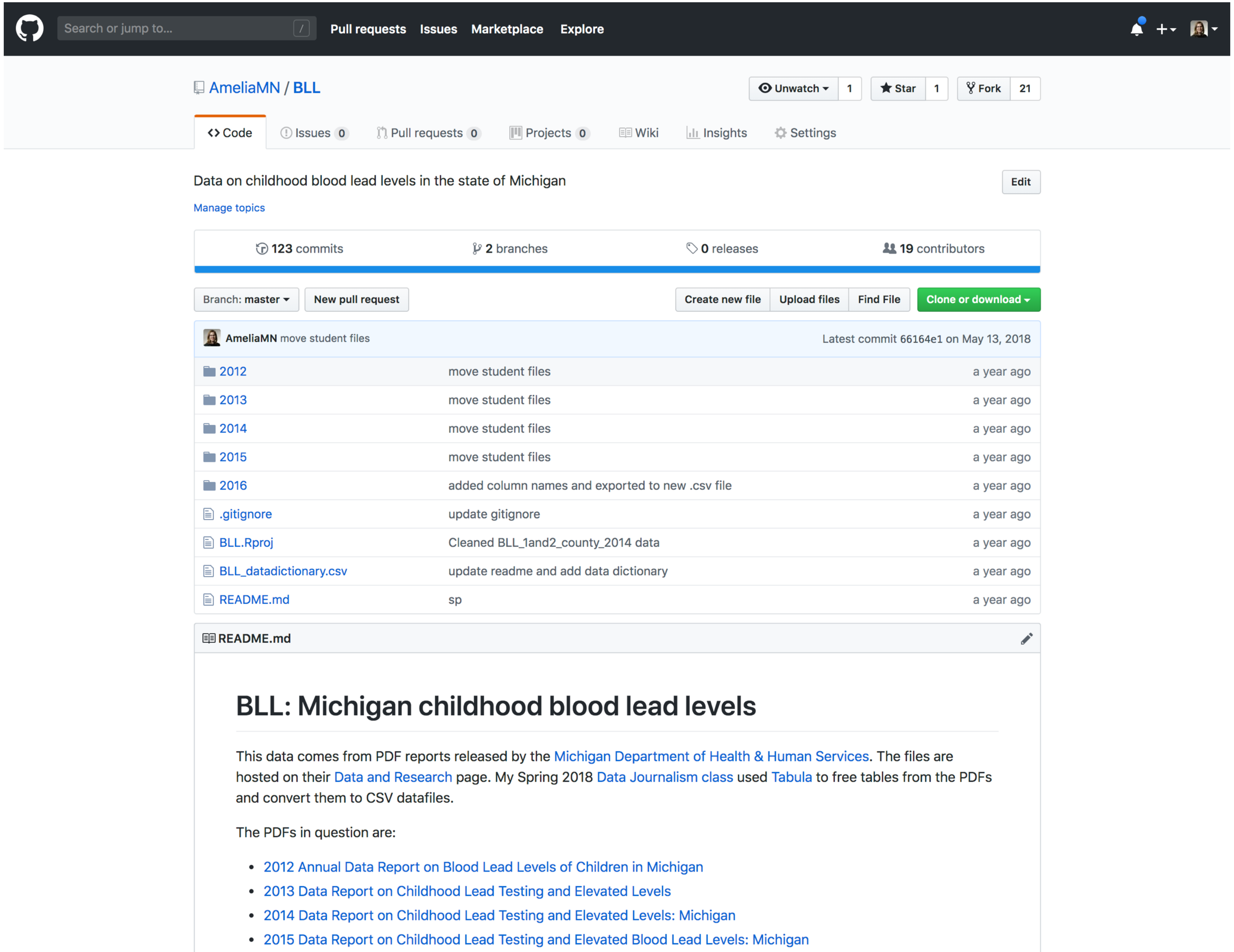
Note: Tabula only works on text-based PDFs, not scanned documents.

Having trouble with Tabula?

1. **Tabula said "Sorry, your PDF file is image-based" -- what does that mean?** Your PDF does not have any embedded text. It might have been scanned from paper. Tabula is not able to extract any data from image-based PDFs. You can try OCRing the PDF with a tool like Adobe Acrobat Pro (paid), Tesseract, PDFSandwich (Mac/Linux, free) or Lime OCR (Windows, free) and then trying Tabula again.
2. **Some columns of my table are combined. What can I do?** Tabula sometimes uses "streams" of whitespace to recreate your table's structure. If headers span multiple columns, they're probably causing a problem. Try excluding them from your selection (or selecting them separately).
3. **Some columns of my table are combined. And the headers aren't the problem! What else can I do?** Tabula has two extraction methods. It tries to guess which one is right for document, but it's wrong sometimes. Try selecting the other (of "stream" and "lattice"), on the left in extraction mode, to see if that fixes the problem.
4. **Tabula helps, but my extracted data isn't in the layout I want! How can I fix that?** Tabula tries to recreate the table structure of the original document. You can think of Tabula as a data extraction tool rather than a data transformation tool. If you want to clean and transform your exported CSV or TSV, tools such as OpenRefine or a spreadsheet program might be a good place to start.
5. **Tabula's taking too long!** Sorry! Tabula has to do a lot of weird math to reconstruct your table. Tabula's command-line counterpart, tabula-extractor is faster, but a little harder to use. You might give it a try.
6. **I had some other problem!** Sorry! You can report it to us here. Be sure to include your PDF, either as a link or attached to the issue -- or email it to one of the Tabula creators.

If you have several PDFs with the same layout, you can select the appropriate regions once, then save the selections as a Tabula Template from the Select Tables page. If someone has shared a template with you, you can upload it to Tabula at the My Templates page.

🔦 Freeing data from PDFS using Tabula



AmeliaMN / BLL

Data on childhood blood lead levels in the state of Michigan

123 commits · 2 branches · 0 releases · 19 contributors

AmeliaMN move student files — Latest commit 66164e1 on May 13, 2018

2012	move student files	a year ago
2013	move student files	a year ago
2014	move student files	a year ago
2015	move student files	a year ago
2016	added column names and exported to new .csv file	a year ago
.gitignore	update gitignore	a year ago
BLL.Rproj	Cleaned BLL_1and2_county_2014 data	a year ago
BLL_datadictionary.csv	update readme and add data dictionary	a year ago
README.md	sp	a year ago

README.md

BLL: Michigan childhood blood lead levels

This data comes from PDF reports released by the Michigan Department of Health & Human Services. The files are hosted on their Data and Research page. My Spring 2018 Data Journalism class used Tabula to free tables from the PDFs and convert them to CSV datafiles.

The PDFs in question are:

- 2012 Annual Data Report on Blood Lead Levels of Children in Michigan
- 2013 Data Report on Childhood Lead Testing and Elevated Levels
- 2014 Data Report on Childhood Lead Testing and Elevated Levels: Michigan
- 2015 Data Report on Childhood Lead Testing and Elevated Blood Lead Levels: Michigan

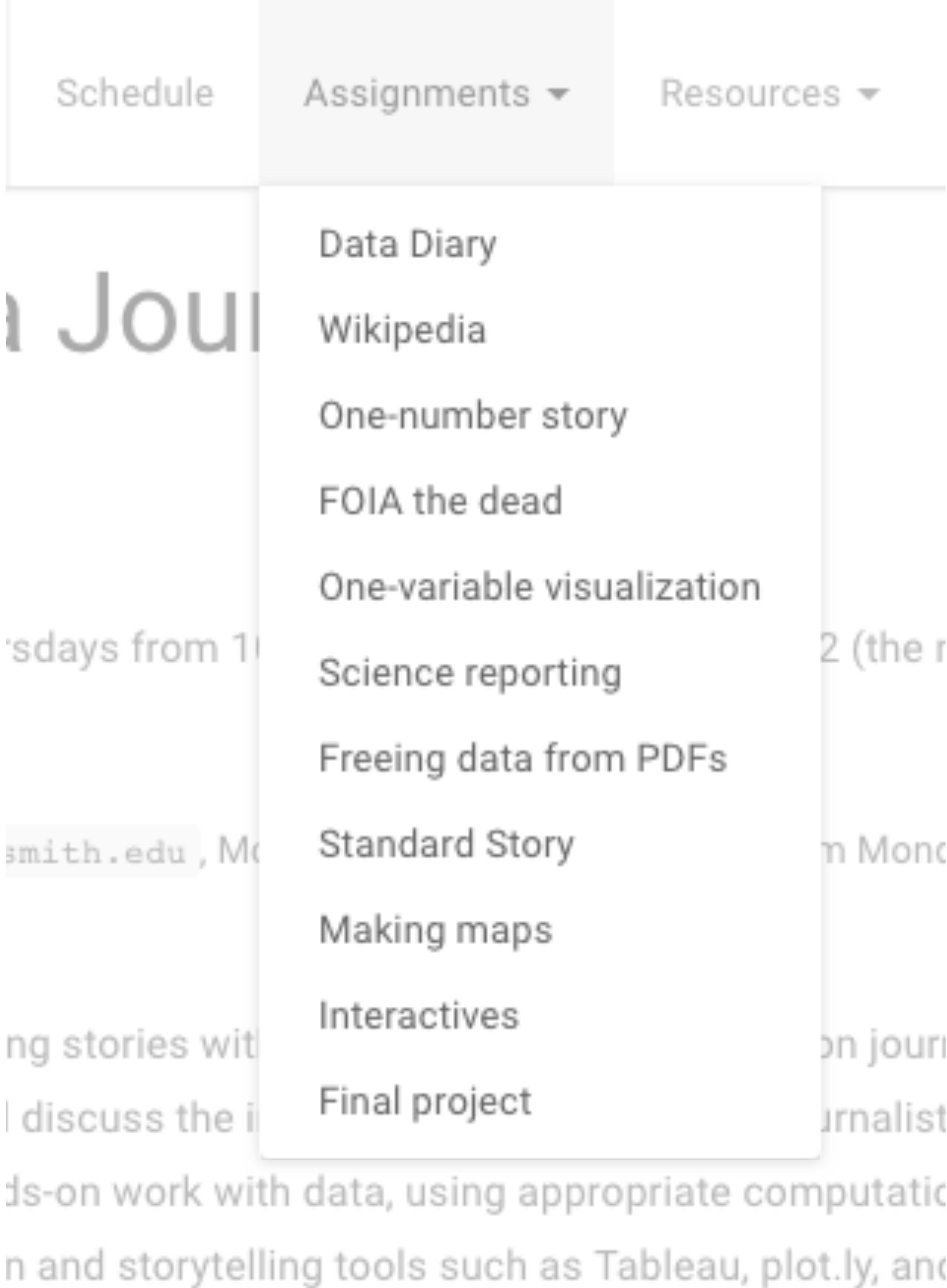
More

journalism

flickr: jannekestaaks

More assignments

- Data diaries
- Authoring Wikipedia articles
- FOIA the dead
- One-variable visualization
- Science reporting
- Standard story
- Making maps



ı Jou

'sdays from 1 2 (the r

smith.edu , Mc n Monc

ng stories wit on jour

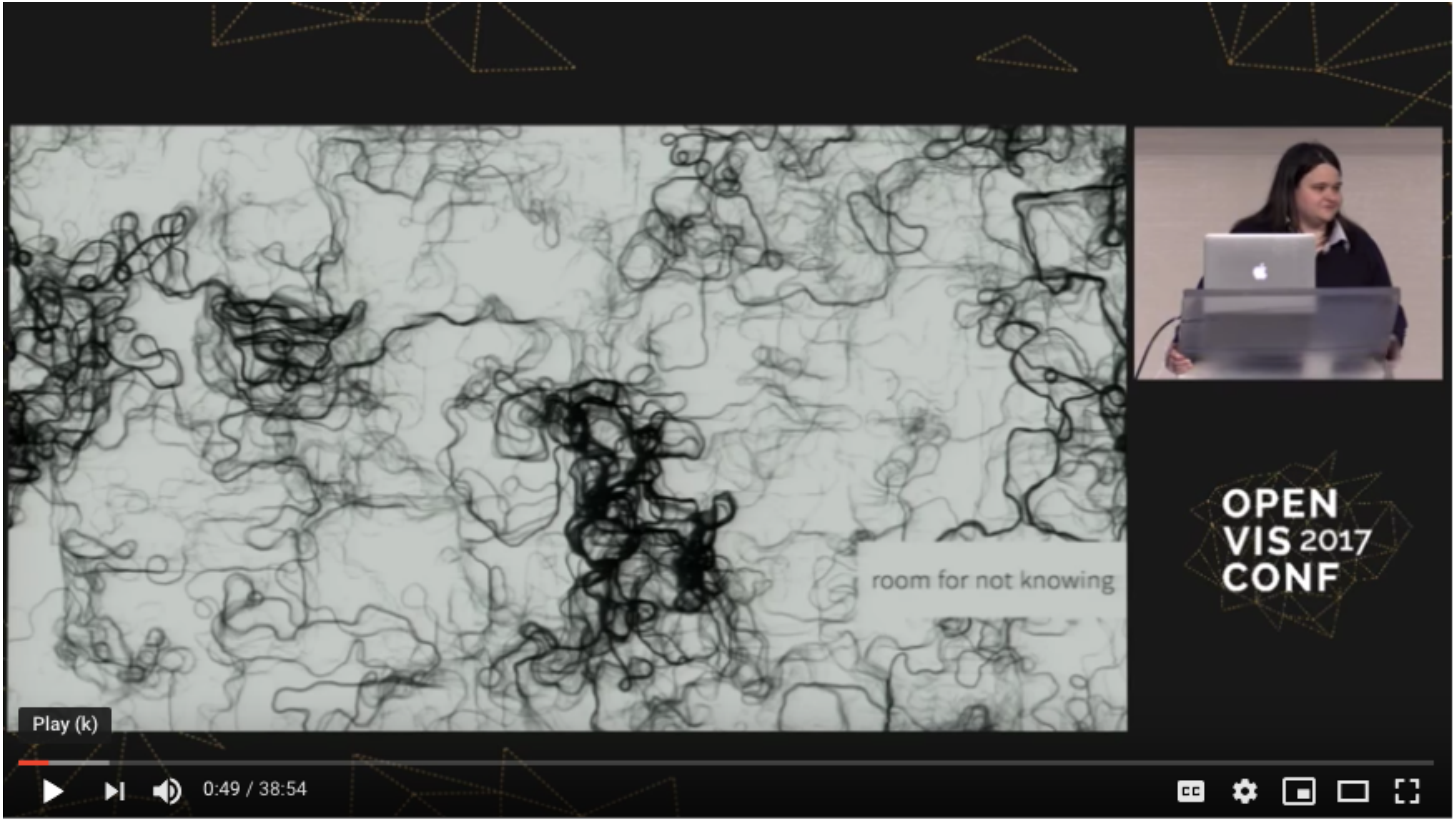
l discuss the i ırnalist

ds-on work with data, using appropriate computatic

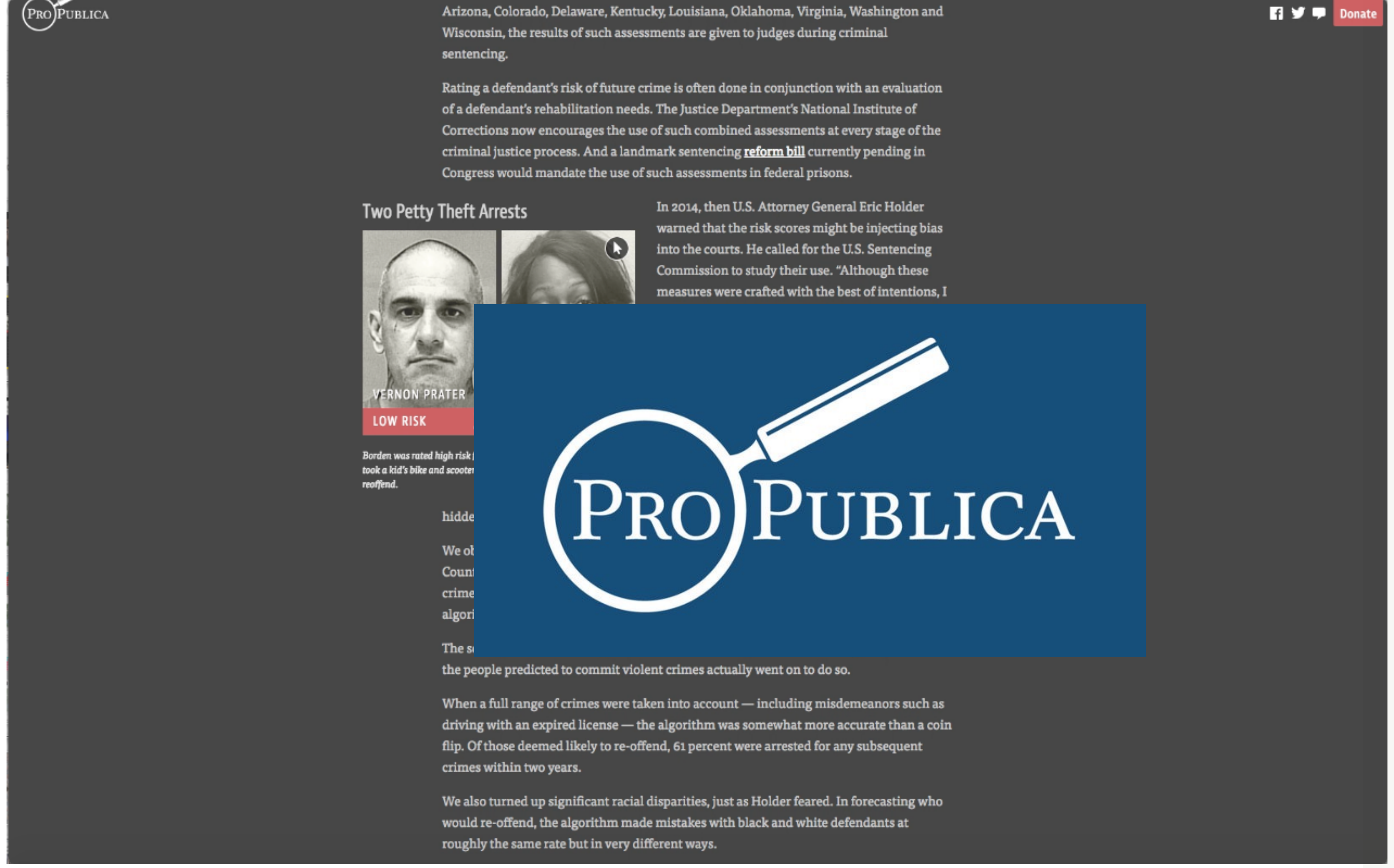
n and storytelling tools such as Tableau, plot.ly, anc

More technology

- Google Drive for organization and spreadsheets
- Slack for communication.
- Tableau for visualization
- OpenRefine for data wrangling
- R and RStudio for data wrangling
- Tabula to free data from PDFs
- git and GitHub to share materials



Amanda Cox, Room for not knowing. OpenVisConf 2017



Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.

Rating a defendant's risk of future crime is often done in conjunction with an evaluation of a defendant's rehabilitation needs. The Justice Department's National Institute of Corrections now encourages the use of such combined assessments at every stage of the criminal justice process. And a landmark sentencing reform bill currently pending in Congress would mandate the use of such assessments in federal prisons.

Two Petty Theft Arrests

Borden was rated high risk f took a kid's bike and scooter reoffend.

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing Commission to study their use. "Although these measures were crafted with the best of intentions, I

hidde

We ob Count crime algori

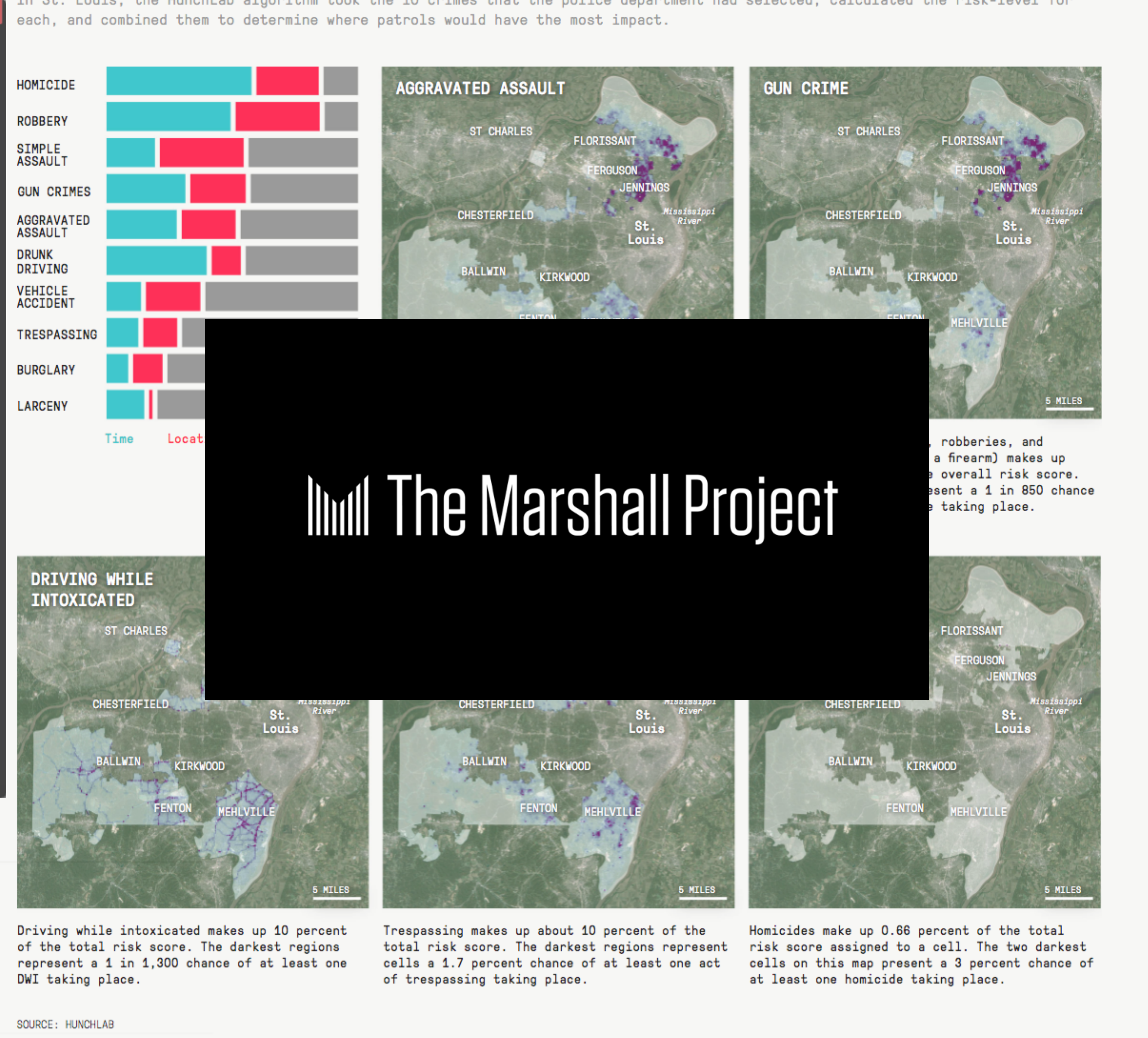
The s

the people predicted to commit violent crimes actually went on to do so.

When a full range of crimes were taken into account — including misdemeanors such as driving with an expired license — the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.

We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

PROPUBLICA



each, and combined them to determine where patrols would have the most impact.

, robberies, and a firearm) makes up e overall risk score. esent a 1 in 850 chance e taking place.

Driving while intoxicated makes up 10 percent of the total risk score. The darkest regions represent a 1 in 1,300 chance of at least one DWI taking place.

Trespassing makes up about 10 percent of the total risk score. The darkest regions represent cells a 1.7 percent chance of at least one act of trespassing taking place.

Homicides make up 0.66 percent of the total risk score assigned to a cell. The two darkest cells on this map present a 3 percent chance of at least one homicide taking place.

SOURCE: HUNCHLAB

The Marshall Project

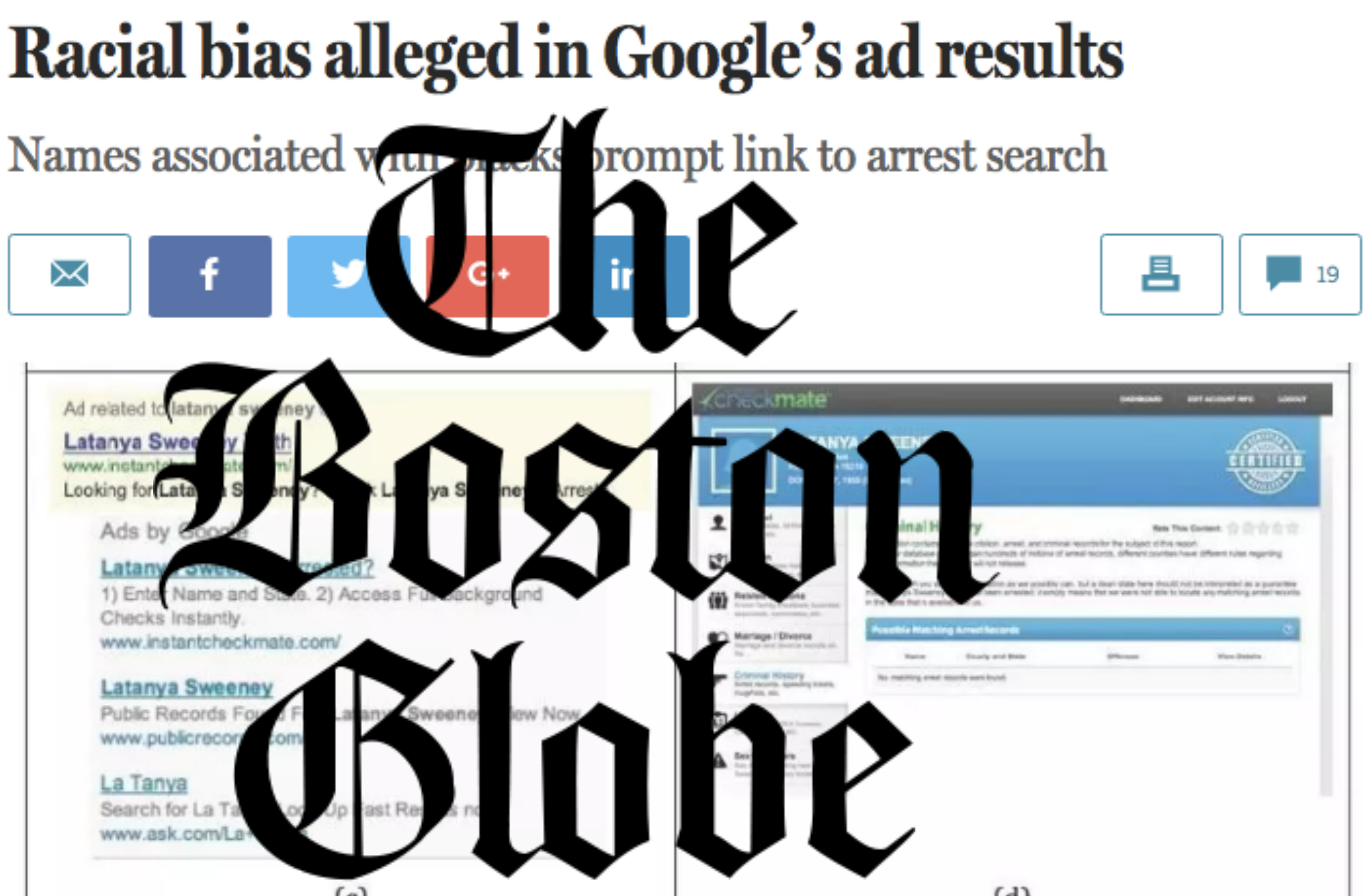
POLITICS

How Trump Consultants Exploited the Facebook Data of Millions

Leer en español

By MATTHEW ROSENBERG, NICHOLAS CONFESSORE and CAROLE CADWALLADR MARCH 17, 2018

The New York Times



Racial bias alleged in Google's ad results

Names associated w[...] prompt link to arrest search

19
Ad related to latanya sweeney
Latanya Sweeney
Looking for Latanya Sweeney?
Ads by Google
Latanya Sweeney Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/
Latanya Sweeney
Public Records Found For Latanya Sweeney. View Now.
www.publicrecords.com
La Tanya
Search for La Tanya Look Up Fast Results now
www.ask.com/La
checkmate
CERTIFIED
Criminal History
Possible Matching Arrest Records

The Boston Globe

IRE/NICAR

Investigative Reporters and Editors is a professional society for journalists, particularly those working in "computer assisted reporting."

Cheap to join ($70/year)

The IRE Journal is a fantastic publication with a behind-the-scenes look at data journalism.

The NICAR conference is a great way to meet data journalists (March 2020, New Orleans).

The IRE Journal
THIRD QUARTER 2017
A SECOND SPOTLIGHT
THE BOSTON GLOBE PUTS SEXUAL ABUSE BACK UNDER THE MICROSCOPE
P. 6
DIGGING INTO 'DEATH CARE'
HOW TO LOOK AT UNFAIR FUNERAL BUSINESS PRACTICES IN YOUR COMMUNITY
P. 14
FOI FILES
AGENCIES STING RECORDS REQUESTERS THROUGH 'REVERSE FOIAS'
P. 34
Watchdog 101:
Scrutinizing colleges and universities
STEPPING OUTSIDE
SECOND

Thank you

Amelia McNamara @AmeliaMN
University of St Thomas
www.amelia.mn

flickr: MattHagedorn